Understanding Traffic Collision Severity's Contributing Factors: A Mixed Effect Multinomial Logistic Regression and Machine Learning Approaches

A Research study submitted to the University Canada West

In Partial Fulfillment of the Requirements

For the Degree of Master's in Business Administration

University Canada West

By

Kawal Walia [redacted]

Abstract

This study aims to understand the influence of various contributing factors on traffic collision severity. With a focus on variables such as pedestrian involvement, cyclist presence, motor vehicle roles, weather conditions, road characteristics, geographical contexts, and among others. The objective of this study is to shed light on the in-depth behavioral dynamics that underlie the severity of accidents. The dataset utilized in this study is retrieved for the Virginia Road department and contains over 500,00 data points with 18 different variables. This study utilized two statistical models and one machine learning model—Multinomial Logistic Regression, Multi-level (Mixed Effect) Multinomial Logistic Regression, which captures the group level heterogeneity, and Random Forest model—to analyze and understand the relationship between various factors and collision severity outcomes. The results show that the Multi-Level Multinomial Logistic Regression model overcomes the Multinomial Logistic Regression model. Moreover, the results show that the existence of vulnerable road users, including pedestrians and bikes would likely increase the odds of fatalities. The odds ratios for fatality and major injuries of collisions involving unbelted drivers are higher than 10, raveling the higher likelihood of sever outcomes compared to belted drivers. Collision occurs are traffic controls (e.g., signalized intersections) are likely to be more severe compared to collision occurs at regular road. These results were in alignment with what were reveled from the Random Forest model. Overall, these findings can help policymakers to design strategies that can reduce severity outcomes in different regions.

Preface

The thesis chapters are submitted for publication in a reputable journal as follows:

Portions of the introductory text in Chapter 1, portions of the literature review in chapter 2,the proportion of the methodology in chapter 3, and the statistical results in chapter 4 are under review in a reputable journal:

Walia, Kawal, & Alsaleh, R. (2023). Exploring the Contributing Factors of Traffic Collision Severity: A Multilevel Multinomial Logistic Regression Model. Submitted.

Contents

1. Introduction

1.1 Background

Car crashes bring about immense human suffering and economic burdens on a global scale. In Canada, traffic collisions are a major contributor to avoidable fatalities, injuries, and about $37 billion in economic losses each year (Transport Canada, 2018). Crafting precise plans to decrease the occurrence, severity, and financial impact of these accidents is a top concern for decision-makers and invested parties. A deeper comprehension of how human actions contribute to these collisions is pivotal in achieving this objective. Although traffic incidents disproportionately affect lower-income and middle-income nations, causing roughly 1.35 million deaths and up to 50 million injuries each year (WHO, 2021), they are also a notable challenge in Canada. In Canada, over 1,800 people face fatal outcomes, and about 160,000 endure injuries yearly due to traffic-related incidents (Transport Canada, 2018). The monetary expenses associated with these accidents in Canada are estimated at around $37 billion annually, a figure that considerably surpasses the human toll. However, delving into specifics, Transport Canada's National Collision Database records 156,310 reported vehicle collisions in 2019. In the corpus of incidents, a total of 1,181 culminated in fatalities, 9,558 in instances of grave injuries, and 137,571 in cases of minor injuries, as reported by Transport Canada in the year 2021. A multitude of inquiries have pinpointed distinct conduct exhibited by drivers that contribute to the probability and intensity of vehicular mishaps. For example, Pang et al. (2019) determined that engaged driving, encompassing activities like operating a mobile device while steering, emerges as a substantial harbinger of road accidents within the confines of Canada, especially prevalent among the demographic of youthful operators. Similarly, Shrestha et al. (2017) established that inebriated driving constitutes a chief instigator of road-related mortalities across Canada, constituting 34% of such fatalities and engendering an annual expenditure of $20.62 billion. Additionally, in consonance with data provided by the Canadian Institute of Actuaries, the fiscal ramifications of motor vehicle collisions amassed to $29 billion in 2018, signifying 1.5% of Canada's gross domestic product (GDP) (Canadian Institute of Actuaries, 2019). This encompassed a gamut of costs, ranging from medical outlays and property destruction to forfeited output and legal charges.

Chauffeur demeanor stands as a solitary component within a nexus of variables that contribute to traffic mishaps, encompassing factors such as road configuration and characteristics of automobiles. Hamer et al. (2021) embarked on an inquiry into the interrelation between Canadian road structure and the propensity for traffic accidents, revealing that road layout constituents like restricted speed thresholds and the incorporation of circular intersections are conducive to mitigating the frequency of vehicular collisions. Correspondingly, Zhu et al. (2018) scrutinized the way diverse vehicular attributes impinge on the gravity of traffic accidents, ascertaining that aged and diminutive vehicles bear an augmented probability of culminating in severe collisions.

Taken together, these investigations accentuate the pivotal function ascribed to driver conduct in dictating the recurrence and severity of road mishaps, alongside the accompanying economic toll, both worldwide and in the context of Canada. Through the cultivation of an augmented comprehension of the precise proclivities that augment the chance and magnitude of road accidents, policymakers and stakeholders are poised to formulate targeted stratagems aimed at abating their incidence and repercussions. Such endeavors, in turn, stand to fashion a transportation framework characterized by enhanced safety and efficiency, redounding to an ameliorated standard of living for the populace of Canada.

1.2 Problem Statement

In terms of road safety and accident prediction, the present research focuses on to work into the intricate interplay of contributing factors and their impact on collision severity. With a detailed and in-depth exploration of all relevant variables including the involvement of pedestrians, cyclists, and motor vehicles, alongside meteorological conditions, road attributes, and area characteristics, this study aims to find the behavioral aspects that contribute to varying levels of accident severity. Central to the research are three distinct predictive models employed to ascertain the severity of collisions: Multinomial Logistic Regression, Multi-Linear Multinomial Logistic Regression, and Random Forest. These models are built to uncover the underlying dynamics governing collision outcomes. The uniqueness of each model lies in their ability to consider the complexity of multiple variables simultaneously and derive comprehensive insights. The focal point of this study is to assess the effectiveness of these models in accurately predicting collision severity, surpassing the current conventional methodologies. By simulating different variables within the Random Forest model, the illumination of the inadequacies of existing models can be noticed, demonstrating that traditional approaches may not capture the nuances of collision severity with precision. The intended outcome of this research is to present an evidence-based framework that empowers road safety agencies, policymakers, and traffic management authorities to adopt proactive measures that align with behavioral patterns. In turn, this can lead to targeted interventions that mitigate accident severity, ultimately fostering a safer and more resilient road environment. The findings of this study not only enrich the understanding of collision dynamics but also pave the way for transformative advancements in road safety strategies.

1.3 Objective

This research study aims to unravel the intricate interplay between human behavior and traffic collision severities. The investigation aims to dissect the multifaceted components inherent in the behavioral dimensions of vehicular collisions, encompassing facets like diverted driving, roadway description, alcohol consumption, excessive speed, and other audacious conduct. Furthermore, the research seeks to disentangle the fiscal consequences associated with road accidents, spanning medical outlays, asset impairment, and compromised productivity.

The core idea of this research study is to explicate the sway of behavioral proclivities on the gravity of road accidents. The inquiry aspires to probe diverse constituents that constitute the behavioral facet of vehicular accidents, encapsulating distracted driving, driving under the influence, speeding, and allied recklessness.

To achieve these aims, this research will utilize statistical and machine learning models that prognosticate the severity of road collisions with different categories grounded in behavioral constituents. These models will scrutinize the efficacy of an array of determinants germane to vehicular mishaps, including driver demeanor, road layout, and vehicle type, in foretelling the magnitude of accidents. The inquiry will lean on trustworthy data sources like law enforcement reports, medical archives, and insurance claims to uphold the precision and dependability of the models.

Preceding research studies have underscored the salience of driver deportment in road accidents, along with their concomitant economic ramifications. For instance, Pang et al. (2019) expounded upon the pivotal role of diverted driving in prognosticating traffic collisions in Canada, while Shrestha et al. (2017) posited that 34% of vehicular fatalities in Canada could be ascribed to drunk driving.

The implications of this research study extend to policymakers, stakeholders, and the general populace, facilitating the formulation of strategies that efficaciously mitigate the recurrence, severity, and economic import of road accidents. By identifying the precise behavioral constituents that contribute to road mishaps, this inquiry can provide insights for precisely calibrated interventions, thereby enhancing road safety and abating the fiscal burden stemming from such incidents. Furthermore, the machine learning models engendered by this study possess the potential to presage the magnitude of road collisions, bolstering the efficiency and efficacy of emergency retorts to such occurrences.

1.4 Significance

The investigation of the role that human behavior plays in determining the severities of traffic collisions constitutes a pivotal realm of research. This domain bears direct implications for both public safety and the economic welfare of a society. The present study delves into the underlying factors contributing to traffic collisions and provides valuable insights into the contributing factors of their severities. These insights, in turn, hold promise for the formulation of efficacious strategies geared towards curtailing the frequency, intensity, and economic aftermath of such collisions.

The crux of this study's importance resides in its innovative harnessing of machine learning models to prognosticate the severity of traffic accidents hinging upon behavioral dynamics. By adopting novel approaches, a prospective avenue opens for emergency services to craft swifter and more effective responses to these incidents. This, in effect, carries the potential to attenuate the loss of human lives and the subsequent economic strain that typically accompanies traffic mishaps. Furthermore, the study's reliance on trustworthy data sources – inclusive of police reports, hospital records, and insurance claims – serves as a safeguard, ensuring the accuracy and dependability of the predictive models generated. The outcomes gleaned from this study can thus serve as the bedrock for targeted interventions aimed at augmenting road safety, mitigating the economic encumbrance tied to such incidents, and formulating more efficacious policies designed to avert traffic accidents.

2. Literature Review

This chapter summarizes the work that were done on identifying traffic collision severities contributing factors. Several previous studies investigated the relation between driving behavior on one side and traffic collision severities and consequence economic impact on another side. For example, Gamage et al. (2021) found a significant correlation between the conduct of drivers and the intensity of vehicular accidents. The repercussions extend beyond mere collisions – the aftermath encompasses escalated instances of accidents, physical injuries, and fatalities, all culminating in a broader economic impact. The ramifications of these incidents manifest as two-fold: direct monetary expenses encompass medical bills, property refurbishments, and legal charges. On a more nuanced level, the indirect costs encompass disruptions to work schedules, compromised quality of life, and elevated insurance premiums (Tlaiss & Baaj, 2020).

The research conducted by Robartes and Chen (2017) investigates the crucial factors influencing the severity of injuries sustained by cyclists in automobile-bicycle crashes using data from Virginia police crash reports spanning the period from 2010 to 2014. Employing an ordered probit model, the study analyzes various crash characteristics, encompassing those related to bicyclists, automobile drivers, vehicles, environmental conditions, and roadways. Their findings reveal significant determinants of injury severity, with intoxicated automobile drivers demonstrating a six-fold increase in the likelihood of cyclist fatalities and double the risk of severe injuries. Bicyclist intoxication also raises the probability of fatalities by 36.7% and doubles the likelihood of severe injury. Moreover, factors such as vehicle speed, obscured driver vision, specific vehicle types (SUVs, trucks, and vans), roadway grades, and curves were identified as contributors to more severe cyclist injuries. The research underscores the need for measures to combat biking and driving under the influence, emphasizing the importance of analyzing and enhancing existing legislation, educating on the perils of drunk driving for cyclists, and promoting the separation of bicycles and vehicles on the road. (Robartes & Chen, 2017).

The in-depth investigation conducted by Agyemang, Li, and Wu (2019) spotlights the critical need to factor in behavioral constituents when devising strategies to mitigate the socio-economic toll stemming from road accidents. Their research accentuates the imperative of implementing effective road safety measures that account for the multifaceted role of human behavior within the context of traffic collisions.

Rahimi et al. (2019) embarked on a study aimed at probing into the severity of injuries resulting from solo-truck crashes in a developing nation. The study's primary objective was to pinpoint the underlying factors that play a role in exacerbating injury severity within such incidents. By harnessing a binary logistic regression model, the study dissected data derived from crash reports detailing 1,690 solo-truck accidents that transpired in Iran between 2011 and 2016. The outcomes disclosed noteworthy influences of injury severity encompassing the age of the driver, speed restrictions, truck arrangement, road type, lighting conditions, and weather scenarios. The authors recommended that an amelioration of road and vehicle conditions, alongside more robust driver training initiatives, might serve to diminish the gravity of injuries in solitary-truck collisions within developing nations.

The menace of road traffic mishaps assumes a substantial role in public health concerns worldwide, carrying both human and financial tolls. In response, scholars have devoted their efforts to investigate the principal risk factors that accentuate traffic accidents and the ensuing injury severity. Rovšek, Batista, and Bogunović (2018) did research to ascertain the primary risk factors governing the severity of injuries arising from traffic accidents on Slovenian roads, using a classification tree method devoid of the constraints of parameter assumptions. Data garnered from traffic accident reports spanning the years 2013 to 2015 in Slovenia were brought under the analytical spotlight. The classification tree analysis, eschewing the need for rigid parameters, divulged those pivotal factors in the intensity of traffic accident injuries included the age of the driver, road type, vehicle category, and nature of the collision. The findings of this study emphasized the necessity of strategies enhancing road infrastructure, bolstering vehicular safety attributes, advocating for prudent driving practices, and elevating driver training programs as the prime avenues for nurturing road safety.

The study conducted by Lestina et al. (1991) addresses a critical issue concerning the understanding of behavior in collision severity, specifically focusing on crashes at freeway entrance and exit ramp interchanges. The research identified and analyzed the most common crash types in heavily traveled urban interstate ramps in Northern Virginia, distinguishing between drivers entering and exiting the freeway. They found that run-off-road, rear-end, and sideswipe/cutoff crashes constituted the majority (95%) of incidents. Notably, run-off-road crashes were most associated with exiting, while rear-end and sideswipe/cutoff crashes were prevalent among entering drivers. Factors such as speed, weather conditions, and alcohol were identified as significant contributors to crash severity.

In a distinct by Zhang et al. (2020), the aim was to unravel the underlying forces shaping traffic breaches and calamity gravity in the realm of China. To accomplish this feat, a stockpile of data culled from the official archives of 4,045 mishaps that unfolded in five of China's urban enclaves, betwixt the years 2016 and 2017, was meticulously scrutinized. The upshot of this endeavor unveiled an assemblage of influential facets predisposing to traffic transgressions and the severity of vehicular mishaps. This array encompassed the chauffeur's vintage, gender, scholastic attainment, motoring history, vehicular genre, road classification, atmospheric state, temporal juncture, and the motorist's comportment, spanning from hastened traversal and temerarious motoring to navigating while under the sway of inebriation. On this account, the scholars proffered the notion that augmentation of driver instruction regimens, the tenacious enforcement of traffic statutes, and an amelioration of vehicular safety attributes could collectively serve to attenuate the frequency of traffic contraventions and the gravity of accidents within the Chinese milieu. This inquiry bequeaths perspicacity that might prove instrumental in the formulation of efficacious stratagems aimed at the advancement of road safety in the Chinese domain.

In Vietnam, Nguyen et al. (2013) investigated the fiscal encumbrance wrought by road traffic traumas upon a provincial general infirmary. In the records of this exploration, information picked from the hospital's ledgers concerning patients felled by road traffic casualties in the span between 2009 and 2010 was diligently analyzed. The expedition yielded revelations of a sizable pecuniary load thrust upon Vietnam due to road traffic injuries, with the monetary outlay averaging a lofty US$1,001 for each sufferer. The upshot of this expedition led to the determination that road traffic mishaps exert a considerable fiscal burden upon both individuals and the collective, and thus, a bevy of preemptive measures, spanning from instilling road safety erudition and the rigorous enforcement of traffic decrees to the amplification of infrastructure standards, should be marshaled forthwith to alleviate the fiscal onus spawned by road traffic casualties.

In another study, Jiang et al. (2020) delved into the ramifications of abnormal weather on fatal road accidents. This inquiry hinged on data extracted from the Chinese Statistical Yearbook and the China Meteorological Data Sharing Service System over the stretch of 2006-2015. The study laid bare a remarkable discovery: extreme weather events, like scorching temperatures and heavy rainfall, displayed a strong connection with escalated chances of lethal road mishaps in China. The findings underlined the potential of adapting to climate shifts by bolstering road infrastructure and transportation networks, countering the adverse effects of freak weather on road safety.

Shifting to a separate perspective, Noland and Quddus (2004) did a research study on a finely detailed analysis of road casualties across England, aiming to uncover the traits that typified regions with elevated road mishap rates. Their lens zoomed in on the timeframe between 1999 and 2001, scrutinizing data at the level of Lower Super Output Areas (LSOAs), diminutive geographical divisions. Employing Poisson regression models, the researchers probed the nexus between road casualties and a slew of factors, including road attributes, socioeconomic variables, and vehicle possession. The study yielded a conspicuous outcome: metropolitan pockets grappling with higher poverty levels and lower car ownership, coupled with roadways featuring raised speed limits, bore the brunt of the worst casualty rates. The team's counsel was crystal clear—direct road safety efforts toward these high-risk locales to curtail road mishaps. In essence, the study provides a treasure trove of insights regarding the spatial arrangement of road casualties and the contributory elements at play in England.

Another study was done with the primary objective of comprehensively exploring the existing conditions and safety challenges faced by nonmotorized vehicles, as well as the role of heavy vehicle drivers in road safety and multi-vehicle collisions within the context of Bangladesh. The initial work by Ahsan and Sufian (2019) utilized secondary data sources encompassing literature reviews and road accident statistics to uncover safety issues linked to nonmotorized vehicles in Bangladesh. The outcomes of this inquiry highlighted that bicycles, rickshaws, and pedestrians are more prone to accidents due to inadequate road infrastructure, insufficient safety precautions, and a lack of adherence to safe driving practices.

Similarly, another research study conducted by Anjuman et al. (2021) within Bangladesh employed a mixed-methods approach to investigate the impact of heavy vehicle drivers on road safety and accidents involving multiple vehicles. This involved surveys, and discussions with drivers, passengers, and road safety experts to amass pertinent insights. The findings underscored that the behaviors of heavy vehicle drivers, such as excessive speed, fatigue, and inadequate training, play a pivotal role in causing road accidents in Bangladesh. The study put forth recommendations including stringent regulations, effective driver training initiatives, and awareness campaigns aimed at enhancing road safety awareness.

This literature review elucidates the critical safety challenges confronted by both non-motorized vehicle operators and heavy vehicle drivers in Bangladesh, providing valuable insights into the country's road safety landscape. Bahrololoom, Young, and Logan (2019) employed a random parameter model to delve into the factors influencing fatalities and severe injuries in bicycle accidents across Victoria, Australia. Drawing from data within the Victorian Traffic Accident System spanning 2012 to 2016, the study identified parameters such as the cyclist's age and gender, time of day, vehicle count, geographic area, and vehicle type as significant predictors of accident severity. The study underscores the potential of enhancing infrastructure, road design, educational initiatives, and adherence to traffic regulations to mitigate the severity of bicycle accidents within Victoria.

In an interesting study undertaken by Aziz et al. (2018), the primary aim was to employ a mixed logit model in order to ascertain the factors influencing the severity of pedestrian-vehicle collisions within the environs of New York City. To accomplish this, the authors drew upon data gleaned from the Motor Vehicle Collision database maintained by the New York City Police Department for the years spanning 2010 through 2014. The study findings demonstrated that variables such as age, gender, vehicle type, geographical location, and time of day held significant predictive power in determining the extent of harm arising from pedestrian accidents. Notably, the study contends that by implementing alterations to policies concerning road layout, enhancing law enforcement measures, and offering more effective educational initiatives, the severity of pedestrian-vehicle collisions in New York City could be mitigated.

Simultaneously, Eluru et al. (2013) undertook a great research study to construct a mixed generalized ordered response model, aimed at comprehending the degrees of harm sustained by pedestrians and cyclists in the context of traffic accidents. The research group derived insights from the South East Queensland Travel Survey spanning 2003 to 2007, along with the Queensland Road Crash Database. Their inquiry revealed the influential role played by variables such as age, gender, vehicle type, travel mode, and geographic location in determining the gravity of accidents involving pedestrians and cyclists. The study advocates for policy adjustments with a focus on augmenting infrastructural elements, refining road design, and bolstering educational and awareness campaigns. These measures are proposed to contribute to the reduction of severity in pedestrian and bicycle accidents within Queensland.

Further, Ehsani et al. (2015) conducted a quasi-experimental study to examine the impact of Michigan's prohibition on texting while driving on the incidence of vehicular accidents. The research aimed to gauge the efficacy of this regulation in curbing the frequency of accidents and injuries resulting from driver distraction. For this purpose, data harnessed from the Michigan State Police Crash Reporting System, encompassing crash records from January 1, 2002, to December 31, 2010, were employed. Employing a difference-in-difference analysis, the study contrasted accident occurrences before and after the enforcement of the texting ban. The outcomes showcased a marked reduction in accidents and injuries attributed to distracted driving following the enactment of the ban, particularly among younger drivers. This study offers insights into potential legislative modifications that could foster a decline in distracted driving incidents.

Alghnam et al. (2018) conducted a case-control investigation in Saudi Arabia to explore the association between cell phone usage while driving and severe injuries resulting from accidents. The study's objective was to identify contributing factors to serious traffic injuries, with a specific focus on the role of cell phone use in such incidents. Data were drawn from a hospital trauma register, with cases representing individuals with major traffic-related injuries and controls encompassing those with minor injuries. Structured interviews were employed to gather data, while logistic regression analysis was utilized to ascertain risk factors. The findings indicated a correlation between talking on a cell phone while driving and an elevated likelihood of experiencing severe car accidents.

Rogeberg & Elvik (2016) employed meta-analysis to investigate the impact of cannabis intoxication on car accidents and subsequent adjustments in perception. The study aimed to ascertain the effects of cannabis use on the occurrence of car accidents and to identify variables influencing this relationship. Prior research on the linkage between marijuana use and vehicular accidents was examined through a meta-analytical approach involving nine studies. The results revealed a heightened probability of car accidents associated with cannabis use, albeit with a smaller effect size than previously believed. The study also determined that the strength of the link between cannabis use and car accidents was influenced by factors such as dosage, frequency, and method of consumption.

Paleti et al. (2010) conducted research with the objective of investigating the influence of aggressive driving behavior on the severity of injuries sustained by drivers in car accidents. Utilizing a multinomial logit model, the study analyzed data from the National Automotive Sampling System Crashworthiness Data System (NASS-CDS) spanning 2004 to 2006. Despite controlling for socioeconomic and crash-related variables, the study established that drivers displaying aggressive behavior were more prone to experiencing severe injuries in collision scenarios.

An insightful study by Wang et al. (2013) aimed to comprehend the impact of traffic dynamics and road characteristics on road safety, subsequently guiding future research directions. Employing a systematic approach, the authors reviewed studies published between 1990 and 2010. The study highlighted the substantial influence of factors such as road configuration, speed limits, and traffic volume on road safety. The authors recommended future investigations center on integrated models that elucidate the intricate interplay between these various factors.

Zhang et al. (2014) delved into the determinants of culpability and injury severity in pedestrian-car accidents within the context of China. Drawing from the Shanghai Traffic Police Department's database, the study employed a binary logit model for analysis. The study unearthed that the pedestrian's age, gender, and crossing location, alongside the driver's age and vehicle type, were pivotal factors in attributing fault and gauging injury severity in such accidents.

Musa et al. (2017) conducted an interesting study into the influence of Malaysia's governmental road conditions on the severity of accidents. Utilizing a binary logit model, the study examined data from the MIROS database, maintained by the Malaysian Institute of Road Safety Research. The investigation underscored the substantial impact of factors such as road geometry, lighting, and road surface conditions on the severity of accidents occurring on federal roads in Malaysia.

Safaei et al. (2021) devised a novel approach employing fuzzy TOPSIS and AHP to prioritize strategies for mitigating motorcycle-related injuries in Iran. They gathered expert insights through a questionnaire and subsequently employed fuzzy TOPSIS and AHP methodologies for analysis. The outcome encompassed a ranked catalog of pivotal criteria and strategies, delineating their significance in curbing motorcycle-linked injuries.

Safaei et al. (2020) aimed to assess the causes and hazards precipitating motorbike accidents in Iran, with a focus on enhancing public safety and well-being. Employing the Delphi method and analytic hierarchy process (AHP), the researchers sequenced selected strategies. Noteworthy findings endorsed enhancing road infrastructure, bolstering awareness and education, and enforcing traffic regulations as optimal measures to reduce motorbike accidents in Iran.

Jacobsen (2013) probed the interconnection between fuel efficiency and safety, delineating its dynamics contingent upon vehicle type and driver behavior. Analysis of data from the Fatality Analysis Reporting System (FARS) and the National Household Travel Survey (NHTS) facilitated the author's estimation of the influence of fuel efficiency on accident rates. Incongruent with conventional wisdom, improved fuel economy did not uniformly heighten accident probability, with the relationship between mileage and safety contingent upon vehicle type and driver conduct.

Zhou et al. (2020) did a study aimed at comparing the severity of public bus accidents arising from collisions versus non-collision incidents. Leveraging data covering bus crashes and related events in Hong Kong spanning 2013 to 2018, the authors conducted analyses employing descriptive statistics and logistic regression. Outcomes indicated that collision-based accidents exhibited graver consequences than their non-collision counterparts. Irrespective of accident type, head, and neck injuries prevailed, advocating the adoption of measures such as seat belts and safety glasses to mitigate the severity of public bus accidents.

Zahabi et al. (2010) sought to gauge the influence of speed limits, built environment, and other factors on the severity of injuries sustained by pedestrians and cyclists in crashes. Drawing from crash data in Montreal, Canada, reported to law enforcement from 2003 to 2013, the researchers employed logistic regression models to explore determinants of injury severity. Results underscored the role of streetlights and pavements in mitigating injuries, while higher vehicular speeds and larger vehicles correlated with more severe outcomes. The study imparts insights into enhancing street safety and decelerating vehicle speeds to ameliorate injury severity among pedestrians and cyclists in accidents.

Wali et al. (2020) scrutinized the relation between fault attribution and injury severity in head-on collisions. Harnessing data from the National Automotive Sampling System Crashworthiness Data System (NASS-CDS) concerning head-on collisions within the United States between 1994 and 2014, the authors employed bivariate ordinal models with copulas to explore the relationship between fault allocation and injury extent. Notably, driving behaviors, encompassing excessive speed and failure to use seat belts, correlated with heightened injury severity. Consequently, curtailing fault allocation emerges as a viable avenue for reducing injury severity in head-on collisions.

Quddus et al. (2015) conducted a study investigating the impact of drivers' geodemographic characteristics on their vulnerability to injury in car accidents. The research employed data from the STATS19 database, encompassing crashes occurring in Great Britain from 2005 to 2009. The scholars employed multilevel mixed effects ordered logit models to explore the influence of various factors on injury severity. The findings demonstrated that older drivers and females exhibited reduced likelihood of sustaining severe injuries. Conversely, drivers in urban areas and those traveling at higher speeds displayed an increased probability of experiencing serious harm. The researchers advocate for strategies such as enhancing road safety measures and reducing vehicle speeds to mitigate injury severity in car accidents.

In a separate study, Osman et al. (2019) examined the extent of injuries sustained by individuals involved in accidents with large trucks within construction zones. The investigation utilized data from the Fatality Analysis Reporting System (FARS) provided by the National Highway Traffic Safety Administration, covering crashes transpiring in the United States from 2007 to 2016. Employing descriptive statistics and ordered logistic regression models, the authors scrutinized factors influencing injury severity. The study disclosed that speed, driver distractions, and the type of truck wielded substantial influence on the seriousness of injuries sustained. The researchers propose interventions like reducing speed limits, bolstering safety measures in construction zones, and enhancing driver awareness to mitigate injury severity resulting from accidents involving large trucks.

Pour et al. (2017) undertook a study with the aim of uncovering the association between neighborhood attributes and the severity of car-pedestrian accidents. The investigation drew upon data from Melbourne, Australia, and employed logistic regression models to analyze how land usage, socioeconomic status, and road characteristics impacted crash severity. The outcomes indicated a heightened likelihood of severe car-pedestrian accidents in areas with increased commercial land use and diminished residential land use. This study underscores the significance of considering the distinctive features of neighborhoods when devising road safety strategies.

Hammad et al. (2019) conducted a study aiming to uncover the impact of environmental factors on car crash occurrences within a suburban region of Pakistan. Drawing upon police records and employing statistical analysis, the researchers explored how variables such as road condition, traffic volume, and weather conditions influence accident rates. The findings revealed a higher frequency of accidents during rainy conditions, on narrow roads, and during periods of heavy traffic. The study underscores the importance of integrating natural elements into road safety planning (Hammad et al., 2019).

Azimian et al. (2020) did a study on an investigation centered around the correlation between area-level attributes and the frequency and severity of automobile accidents. Using crash data from Houston, Texas, the researchers employed a multivariate space-time model to ascertain the impact of factors like land use, socioeconomic status, and road attributes on crash incidents. The study demonstrated a heightened likelihood of severe crashes in regions dominated by business land use and reduced residential land use. The implications suggest that altering land usage and refining transportation planning could contribute to a reduction in both the number and severity of traffic accidents.

Pasha et al. (2016) did a study into the interplay between street layout, traffic flow, road infrastructure, socioeconomic elements, and demographic factors influencing public transportation utilization. Based on data from Dhaka, Bangladesh, the researchers employed regression models to dissect the effects of diverse variables on traffic patterns. The outcomes spotlight the substantial influence of factors such as population density, road width, and traffic volume on public transport utilization. The study underscores the necessity of comprehensive considerations in public transport planning, offering valuable insights to policymakers in developing nations.

Sapkota, Bista, and Adhikari (2021) aimed to quantify the economic repercussions of motorbike accidents in Kathmandu, Nepal. Utilizing data from the Government of Nepal's Metropolitan Traffic Police Division spanning from 2013 to 2017, the experts assessed the financial impact of motorbike crashes, encompassing direct costs like medical expenses and property damage, as well as indirect costs like lost productivity. The study unveiled a staggering total cost of NPR 8.25 billion (equivalent to USD 72.16 million) to the economy during the study period. The findings underline the urgency for policy adjustments to enhance road safety and mitigate the economic burden of motorbike accidents in Nepal.

3. Methodology

3.1 Introduction

As part of our study, statistical and machine-learning models will be developed to foresee Collision Severity and determine their contributing factors. In this section, we'll delve into statistical models: the multinomial logistic regression model and the multi-level multinomial logistic regression model. The multinomial logistic regression model will help us scrutinize crash severities by considering various behavioral factors and forecasting the probabilities of diverse severity groups. The multi-level multinomial logistic regression model will boost our analysis by considering the data's hierarchical setup and incorporating predictors at both individual and group levels. With these approaches, our objective is to uncover pivotal elements influencing crash severity and their ramifications.

3.2 Data Description

This chapter will investigate the data that was used in research study focused on predicting collision severity. The data was obtained from the Virginia Road Department, where it's meticulously collected based on reported collisions. Notably, the data is accessible to the public and promotes transparency.

The dataset used represents a comprehensive compilation of collisions transpiring between 2019 and 2023. These collisions were promptly reported to the Virginia Road Department. The dataset consisting of approximately 500,000 collision incidents. The dataset contains both categorical and numerical attributes. These attributes furnish us with diverse insights that are crucial for our prediction model.

3.2.1 Data Processing Steps

To ensure the reliability and accuracy of our model, a study undertook a series of meticulous data processing steps. These steps encompassed cleaning, transformation, and feature extraction. The critical phase of data cleaning is started first. This cleansing process bolstered the quality of our data, ensuring that our model's foundation is sturdy and dependable. Given the mix of categorical and numerical attributes, the power of encoding techniques was utilized to transform categorical variables into numerical representations. This pivotal step enabled our model to understand and learn from the data more effectively. By translating categorical attributes into numerical factors, the groundwork for comprehensive analysis was laid.

Delving deeper, feature extraction was engaged in. This intricate process fortified our model's potential for generalization, noise reduction, and overall stability. It enabled our model to focus on the most impactful attributes while diminishing the influence of extraneous factors. This, in turn, heightened the precision of our predictions.

As the aim of this research is predicting collision severity, each of these data-related strides was pivotal. The rich data, coupled with meticulous processing, fortified the foundation of our research study. By refining the dataset to recent years, it was ensured that the model remains attuned to contemporary patterns. The transformation of categorical attributes into numerical forms heightened our model's perceptiveness, and feature extraction optimized its predictive capabilities.

3.2 Variables

To achieve the objective of forecasting collision severity accurately, the selection of predictor variables plays a pivotal role. This section delves into the process of picking the right variables to empower our prediction model with the ability to discern and anticipate the severity of collisions.

When this research project was initiated, a set of potential predictor variables was assembled by drawing on both domain knowledge and existing research. These variables, it was surmised, might wield the influence to convert the complex and big tapestry of factors that contribute to collision severity. The culmination of this phase brought forth a promising array of attributes ready to be scrutinized for their predictive prowess.

A vital step on this road to precision was conducting an exploratory data analysis (EDA). This analytical journey unfurled insights into the relationships between variables and their role in shaping collision severity. Imagine it peering through a magnifying glass at the complex web of data. Our focus was on understanding how each variable interacts with the target variable – collision severity.

Visualizing the correlation matrix emerged as an invaluable tool during this phase. This matrix unveiled a visual relation of interconnectedness, allowing us to perceive which variables danced in harmony and which ones stood out in stark contrast. These visualizations provided a tangible glimpse into the complex relationships that underlie collision severity, and the most influential or important factors are mentioned in Table 3.1.

Table 3.1
Variables in our dataset to do the modeling.

Variables	Description	
VDOT_DISTRICT	Accidents for districts in Virginia	
WEATHER CONDITION	Collision in Rainy condition	16% (Adverse)
ROADWAY ALLIGNMENT	Collision at different alignment like straight or intersection	14.71% (Curve) 85% (Straight)
ROADWAY DESCRIPTION	Collision based on road type	3.19% (One-Way) 57% (2-way divided) 39.73% (2-way undivided)
COLLISION TYPE	In what way did the collision happened	26.22 % (Angle) 21.7 % (Fixed) 2.3% (Head On) 1.6% (No Collision) 27.51 % (Rear- End) 10.04 % (Side swipe)
ALCOHOL	Driver under influence of alcohol at the time of collision	5.8%
UNBELTED	At the time of Collision if the seat belts were on or not	4.6%
BIKE	Collision with Bike or Not	0.48%
TRAFFIC SIGNAL	If the traffic control signal were present in the area	80.4 %
AREA TYPE (Urban)	If the collision happened in Urban or Rural area	74.27% (Urban) 25.77% (Rural)
DISTRACTED	Any distractions in external environment for driver	17.43%

DROWSY	Sleepy	2.7%
CRASH DATE	Collison happened on weekday or weekend	73.61% (Weekday) 26.38 % (Weekend)
DRUG	Driver under influence of drugs/medicine at the time of collision	1.01%
MOTOR	Were any Car or motor vehicle involved in the collision	1.71%
PED	Any Pedestrian involved	1.18%
SPEED	Driver speeding or not at the time of collision	20.77%
ANIMAL	Collision involving Animals	6.4%

3.3 Multinomial Logistic Regression Model

3.3.1 Description

The multinomial logistic regression model is a method used to analyze and predict results with more than two categories. In the realm of traffic accidents, this model lets us comprehend the human behavioral, environmental, and other elements that influence crash severity and their economic repercussions. This methodology will lay out the steps to apply the multinomial logistic regression model in our research, which focuses on comprehending the impact of behavioral facets on traffic collision seriousness and its economic implications.

When this research project was initiated, a set of potential predictor variables was assembled by drawing on both domain knowledge and existing research. These variables were surmised to potentially wield influence in converting the complex and substantial tapestry of factors contributing to collision severity. A promising array of attributes ready to be scrutinized for their predictive prowess was brought forth by the culmination of this phase.

A vital step on this road to precision was taken as an exploratory data analysis (EDA) was conducted. Insights into the relationships between variables and their role in shaping collision severity were unfurled through this analytical journey. The complex web of data was examined as if through a magnifying glass. Emphasis was placed on understanding how each variable interacts with the target variable – collision severityThis analysis will provide insights into the key behavioral aspects that need to be addressed to mitigate collision severity.

3.3.2 Equation

The multinomial logistic regression model estimates the log odds of each crash severity category relative to a reference category. The general equation for the multinomial logistic regression model can be expressed as follows:

$$\log (P(Y = j \mid X)) = \beta j0 + \beta j1X1 + \beta j2X2 + ... + \beta jkXk,$$

where:

- $P(Y = j \mid X)$ represents the probability of crash severity category j given predictor variables X.

- βj0, βj1, βj2, ..., βjk represent the estimated coefficients for each predictor variable in the model.
- X1, X2, ..., Xk represent the predictor variables (behavioral aspects, road conditions, etc.) influencing crash severity.

3.3.3 Model Development

Our aim for prediction leads us to the use of one of the relevant models which is Multinomial Logistic Regression. This model, grounded in statistical principles, is designed to predict the degrees of crash severity and is helpful when multiple categories are there in the targeted variable.

In the core of the model development process, the analysis was done to categorize into distinct sections. The dataset was categorized into multiple binary logistic regression sub-models, comparing each severity category against the reference category (e.g., minor severity vs. reference, severe severity vs. reference). This step ensures that our model captures the shades of difference that define each level of severity.

The goal of implementing this approach lies in training multiple binary logistic regression sub-models. Each sub-model stands as a beacon of insight, illuminating the intricate relationships between predictor variables and crash severity. By dissecting the dataset into these sub-models, the data is navigated with precision, enabling to grasp the factors that play a pivotal role in determining the outcome of a collision.

This model development phase is not just a technical interlude; it's the cornerstone of our research study's empirical foundation. By choosing Multinomial Logistic Regression, methodology is aligned with the complexities of real-world collision scenarios.. The categorization and comparison offer a lens to perceive the spectrum of severity levels, much like dissecting light through a prism.

To conclude, the multinomial logistic regression model is a valuable tool for analyzing crash severity in traffic collision studies. By employing this methodology, the behavioral aspects that significantly influence crash severity and their impact on the economy can be identified. The model's estimated coefficients and odds ratios allow for a quantitative understanding of the relationship between predictor variables and crash severity categories. Implementing this methodology will contribute to our research study's objective of comprehending the influence of behavioral aspects in traffic collision severities and its broader economic implications.

3.4 Multi-level Multinomial Logistic Regression

3.4.1 Description

The multi-level multinomial logistic regression model is an extension of the multinomial logistic regression model that allows for the incorporation of hierarchical or nested data structures. In the context of our research study on understanding the influence of behavioral aspects in traffic collision severities and its impact on the economy, this methodology outlines the steps involved in applying the multi-level multinomial logistic regression model to analyze crash severity while accounting for nested data structures.

Furthermore, the variation in crash severity across different locations results in a level of diversity within specific groups that is smaller than the diversity observed between these groups (Tang et al., 2020). This phenomenon becomes evident when considering factors like geographic divisions, differences between rural and urban areas, land use patterns, climate zones, and functional areas (Peng et al., 2019). In essence, the spatial differences play a role in shaping the severity of injuries resulting from crashes. It's of utmost importance to take all levels of clustering into account when conducting an analysis of crash data that involves multiple levels. Failing to consider the effects specific to each cluster could introduce statistical errors, which might manifest as skewed parameter estimates, underestimation of standard errors, and an exaggerated sense of statistical significance.

3.4.2 Equation

The general equation for the multi-level multinomial logistic regression model can be expressed as follows:

$$\log (P(Yij = j \mid Xij)) = \beta0j + \beta1jX1ij + \beta2jX2ij + ... + \beta kjXkij + u0j + u1jW1j + u2jW2j + ... + ukjWkj,$$

where:

- $P(Yij = j \mid Xij)$ represents the probability of crash severity category j for individual i within group j given predictor variables Xij.
- $\beta0j, \beta1j, \beta2j, ..., \beta kj$ represent the fixed effects coefficients for individual-level predictors.
- $X1ij, X2ij, ..., Xkij$ represent the individual-level predictor variables influencing crash severity.
- $u0j, u1j, u2j, ..., ukj$ represent the random effects coefficients for group-level predictors.
- $W1j, W2j, ..., Wkj$ represent the group-level predictor variables influencing crash severity.

3.4.3 Model Development

This model considers the unique structure of our data, accommodating the fact that crashes can vary in severity. In this process, factors that pertain to both individual cases and broader groups are considered. To ensure precision, the dataset is segmented into different categories of crash severity. Each category is compared against a reference point, essentially helping the model understand what distinguishes, for instance, a minor collision from a more severe one.This breakdown adds depth to our analysis, as it helps us capture the nuances inherent in various levels of crash outcomes.

Furthermore, our model isn't just focused on individual cases; it also recognizes that different groups might exhibit distinct patterns. This is where the concept of random effects comes into play. These effects are incorporated to capture the variations that arise between different groups, which might experience crashes differently due to unique circumstances. In the model, V_DOT is going to be kept as the group level. The estimation of random effects at the group level captures the variations across diverse groups, enabling us to analyze the contextual effects on crash severity.

3.4.3 Model Evaluation

After running the model and defining all variables and conditions, the model can be evaluated by assessing the significance and contribution of each predictor at both individual and group levels using statistical tests, such as p-values or Bayesian model comparison. Moreover, to have a better understanding of the results the emphasis can also be given at the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). As these models were built as explanatory models (not predicted), the main aim is to identify the contributing factors in collision severities, rather than predicting the collision severities.

In our research study, the multi-level multinomial logistic regression model allows us to account for the hierarchical structure of the data and examine the influence of individual-level and group-level factors on crash severity. By including individual-level behavioral aspects and group-level factors such V-DOT in our model, it can give a comprehensive understanding of the factors contributing to crash severity. The estimation of random effects at the group level captures the variations across diverse groups, enabling us to analyze the contextual effects on crash severity.

This analysis will provide valuable insights for policymakers, transportation planners, and stakeholders to develop targeted interventions and strategies at the individual and group levels.

3.5 Random Forrest Model

3.5.1 Description

In understanding and predicting collision severity, a comprehensive research study will be conducted, utilizing the power of data analysis and machine learning techniques. In this part of the methodology, the step-by-step process followed to build a Random Forest model aimed at accurately predicting and classifying collision severity using data collected from the North Virginia Federal Department website will be outlined. The approach encompasses data collection, data preprocessing, feature engineering, model building, and evaluation.

3.5.2 Equation

Random Forest consists of a group of classification trees, where each classification tree is trained based on a random sample from the input dataset (Breiman, 1999). The Random Forest utilized in this study used a random number of features to be considered at each split to grow the trees. Each decision tree votes towards the classification decision, and results of the classification is specified using the Gini Index. At a given input T, selecting the Collison severity level and specifying that it belongs to category C_i, a Gini Index can be given by:

$$\sum \sum_{j \neq i} (f(C_i, T)/|T|)(f(C_j, T)/|T|)$$

where f(Ci,T)/|T| is the probability that the selected case belongs to class Ci.

The Random Forest would grow N number of trees and the decision would be for the categories with eh maximum number of votes for the collision severities.

3.5.3 Model Development:

The main part of our methodology lies in the construction of the Random Forest model. A Random Forest is an ensemble learning technique that combines multiple decision trees to make more accurate predictions. The model was trained using the training data, and 100 decision trees were used as our ensemble.. This forest of trees collaboratively worked together to predict collision severity.

3.5.4 Model Evaluation:

The model's true prowess shines in its ability to make accurate predictions. To evaluate its performance, the testing dataset is used that had been kept aside. Accuracy, a widely used classification metric, was employed to measure how well the model's predictions aligned with the actual collision severity labels. This metric provided us with a tangible understanding of the model's accuracy in predicting different severity levels.

4. Statistical Modeling Results

In this chapter, the results of the developed model will be discussed. Three models are developed in this thesis, including (1) multinomial logistic regression model, (2) multi-level multinomial logistic regression model, and (3) Random Forest model. For all three models, the target variable remains the same, which is Collision Severity, where 4 categories are retained: Minor Injury, Major Injury, PDO, and Fatality. The different categories of Collision Severity will be examined using the term KABCO, which assesses and classifies the severity of collisions if they lead to Fatality, Major Injury, Minor Injury, or just property damage. In our modeling, PDO (property damage) was kept as the baseline.

4.1 Multinomial Logistic Regressions

The parameter estimates for the multinomial logistic regression model are presented in Table 4.1.

Table 4.1
Output from MNL Model including Coefficients, p-value and Odds Ratio

Variables bike-vehicle	FATALITY			MAJOR INJURY			MINOR INJURY		
	Coef.	P-Value	Odds Ratio (exp(coef))	Coef.	P-Value	Odds Ratio (exp(coef))	Coef	P-Value	Odds Ratio (exp(coef))
Intercept	-14.9	0		-2.4	<0.01		-1	<0.01	
Collision Type (Head-On) Angle*	-0.02	.07	0.98	-	-	-	-	-	-
Collision Type (Rear End) Angle*	-1.4	0.07	0.19	-0.75	.04	0.50	-	-	-
Traffic Control (Yes) (No)*	10.07	<0.01	4.6	-0.25	<0.01	0.83	-0.08	<0.01	1.15

Traffic Control (Other) (No)*	9.1	<0.01	1.5	-0.58	<0.01	0.59	-0.3	<0.01	0.90
Belted (No) (Yes)*	3.2	<0.01	29.25	2.2	<0.02	10.03	-	-	-
Bike (Yes) (No)*	5.06	<0.01	269.07	4.3	<0.01	104	3.6	<0.01	51
Animal (Yes) (No)*	-10.6	<0.01	0.00	-0.1	<0.01	0.53	-0.1	<0.01	0.8
Pedestrians (Yes) (No)*	1.05	<0.01	2841	1.05	<0.01	4952	1.03	<0.01	2248
AIC	4275949.80422								
BIC	428464.30900								

The equations for the multinomial logistic regression model developed for the crash severity are:

Fatality:

$$\text{Log}(P(Y = \text{Fatal}/\text{PDO})) = -14.9 - 0.02X1 - 1.14X2 + 10.07X3 + 9.1X4 + 3.2X5 + 5.06X6 - 10.06X7 + 1.5\,X8$$

Major Injury:

$$\text{Log}(P(Y = \text{Major\_Inj}/\text{PDO})) = -2.4 - 0.02X1 - 0.75X2 - 0.25X3 - 0.58X4 + 2.2X5 + 4.3X6 - 0.1X7 + 1.05X8$$

Minor Injury:

$$\text{Log}(P(Y = \text{Major\_injury}/\text{PDO})) = -1 - 0.3X1 - 0.2X2 - 0.08X3 - 0.3X4 + 1.3X5 + 3.6X6 - 0.1X7 + 1.03X8$$

where:

X1	Collision Type (Head- On)
X2	Collision Type (Rear End)
X3	Traffic Control (Yes)
X4	Traffic Control (Other)
X5	Belted (No)
X6	Bike (Yes)
X7	Animal (Yes)
X8	Pedestrians (Yes)

After the Multinomial Logistic Regression model was run for the dataset, output including Coefficients, Std error, and P-values of the coefficients was obtained. In this part of the chapter, multiple significant variables will be discussed, and the effect of the variable in predicting collision severity will be concluded. The different categories of Collision Severity will be examined using the term KABCO, which assesses and classifies the severity of collisions if they lead to Fatality, Major Injury, Minor Injury, or just property damage. The outcome will be analyzed using the Odds Ratio. The Odds Ratio can be defined as the ratio of the probability of two events. For our modeling, PDO (property damage) was kept as the baseline. The Odds ratio of different variables for different collision severity can be examined.

Collision Type:

Different collision types, such as Head-on, Rear end, Side sweep, and Fixed, were analyzed with the reference category kept as Angle. The odds ratios associated with collision type 'Head On' provide valuable insights into the impact on injury severity outcomes. Specifically, the odds ratio for fatality is 4.9, indicating a substantially higher likelihood, followed by a 3.4-fold increase for major injuries, and a 1.6-fold increase for minor injuries. These findings underline the significant influence of collision type on injury outcomes compared to the reference category. On the other, the odds ratio linked with Collision type- Rear end for fatality stands at 0.19, implying a notably lower likelihood compared to the reference category. For major injuries, the odds ratio is 0.50, while for minor injuries, it approximates at 1.03. This indicates that Head-on collisions are more severe as compared to other type of collision recorded (see Table 4.1).

Traffic Signal:

Looking at collision outcomes, the presence or absence of traffic control emerges as a pivotal factor. When compared to collisions devoid of traffic control measures, those featuring traffic control display an odds ratio of 4.6 for fatality, indicating a notably elevated likelihood. This underlines the potential protective effect of traffic control measures, which appear to mitigate the risk of minor injuries, as reflected in the odds ratio of 1.15. Interestingly, for major injuries, the odds ratio of 0.83 tells a reduced probability in these scenarios (see Table 4.1).

Traffic Control (Other):

Further looking at our model outcomes, the category 'Traffic Control (Other)' emerges as an intermediary. In comparison to the baseline, 'No' category, collisions falling under this classification exhibit odds ratios of 1.5 for fatality, 0.59 for major injury, and 0.9 for minor injury.

These odds ratios serve as markers, denoting the nuanced impact that 'Traffic Control (Other)' has on each level of injury severity.

Belted:

The fastened seatbelt shows significant influence. In comparison to the baseline 'Yes' category, collisions, where individuals are not belted, show odds ratios of 29.25 for fatality, 10.03 for major injury, and 3.6 for minor injury. These odds ratios unveil a reality – the absence of seatbelt usage dramatically heightens the likelihood of severe outcomes specially Fatality and Major injury. The substantial odds ratios for fatality and major injury underscore the pivotal role seatbelts play in mitigating the risk of grave consequences. The odds ratio for minor injuries, while still elevated, denotes a relatively moderated impact. These findings unequivocally advocate for the universal adoption of seatbelt practices to foster a safer road environment (see Table 4.1).

Bike:

Looking at the involvement of bicycles in collisions, a striking pattern emerges. Collisions involving bikes reveal odds ratios of 269 for fatality, 104 for major injury, and 51 for minor injury, relative to the baseline. This highlights the vulnerability associated with bicycle-involved accidents, where the odds of fatality and major injury are very high. The odds ratio for minor injuries, while still substantial, appears comparatively moderate, shedding light on the relative impact.

Animal:

Table 4.1 gives an insight about the influence of animal involvement in the predicting the severity. The presence of animals, often unpredictable elements, lends a unique perspective. In contrast to the baseline, 'No' category, collisions involving animals display odds ratios of 0 for fatality, 0.53 for major injury, and 0.8 for minor injury. The fatality odds ratio of 0 suggests a potential protective factor, although a cautious interpretation is warranted given the rarity of animal-related fatalities. The odds ratios for major and minor injuries indicate a modestly elevated risk, highlighting the potential for animal-involved collisions to lead to a higher likelihood of injury.

Pedestrians:

Considering the last significant variable of our model, pedestrians introduce a significant dimension. Collisions involving pedestrians present contrasting odds ratios across the severity spectrum. With odds ratios of 28441 for fatality, 4952 for major injury, and 2248 for minor injury compared to the baseline 'No' category, the message is unequivocal: pedestrian-involved collisions pose a dramatically heightened risk across all injury levels. These odds ratios serve as a clarion call for increased attention to pedestrian safety measures.

4.2 Multi-Level Multinomial Logistical Regression

The parameter estimates for the multi-level multinomial logistic regression model are presented in Table 4.2.

Table 4.2
Output from Multi-level MNL Model including Coefficients, p-value and Odds Ratio

Variables bike-vehicle	FATALITY			MAJOR INJURY			MINOR INJURY		
	Coef .	P-Value	Odds Ratio (exp(coef))	Coef .	P-Value	Odds Ratio (exp(coef))	Coef.	P-Value	Odds Ratio (exp(coef))
Intercept	-10.4	<0.01		-7.01	<0.01		-0.7	<0.01	
Collision Type (Head-On) *Angle	1.85	0.06	1.9	1.4	<0.01	3.6	0.62	<0.5	1.2
Collision Type (Rear End) *Angle	-0.34	<0.01	0.75	-0.69	0.04	0.65	-	-	
Collision Type(Side sweep) *Angle	-0.38	<0.01	0.68	-1.1	0.02	0.38	-	-	-
Traffic Control (Yes) (No)*	1.6	<0.01	46	4.1	<0.01	431	-0.64	<0.01	1
Traffic Control (Other) (No)*	0.25	<0.01	18	4.4	<0.01	345	-0.41	<0.01	0.84
Belted (No) (Yes)*	3.8	<0.01	44.70	2.3	<0.01	10	1.2	<0.01	4
Bike (Yes) (No)*	4.5	<0.01	13	3.6	<0.01	67	2.9	<0.01	41
Animal (Yes) (No)*	-0.76	<0.01	5.7	-0.46	<0.01	0.56	-	-	-
Pedestrians (Yes) (No)*	9.2	<0.01	354	7.0	<0.01	436	5.7	<0.01	258
Area Type (Urban) Rural*	-1.6	0.10	0.60	-	-	-	-	-	-

Roadway Description (2-way divide) (One way) *	2.8	<0.03	1.7	0.49	0.06	2.7	-	-	-
Random Effect Coef.:									
Var (V_dot)	0.67			0.64			0.86		
Goodness of fit:									
AIC	421118								
BIC	422365								

The equations for the Multi-level multinomial logistic regression model developed for the crash severity are:

Fatality-

$$\text{Log}(P(Y = \text{Fatal}/\text{PDO})) = -10.4 + 1.85X1 - 0.38X2 - 0.38X3 + 1.6X4 + 0.25X5 + 3.8X6 + 4.5X7 - 0.76X8 + 9.2X9 - 1.6X10 + 2.8X11$$

Major Injury

$$\text{Log}(P(Y = \text{Major\_inj}/\text{PDO})) = -7.01 + 1.4X1 - 0.69X2 - 1.1X3 + 4.1X4 + 4.4X5 + 2.3X6 + 3.6X7 - 0.46X8 + 7\,X9 - 0.46X10 + 0.49X11$$

Minor Injury

$$\text{Log}(P(Y = \text{Minor\_Inj}/\text{PDO})) = -0.7 + 0.62X1 - 0.14X2 - 0.28X3 - 0.64X4 - 0.41X5 + 1.2X6 + 2.9X7 - 0.23X8 + 5.7X9 - 0.24X10 + 0.32X11$$

where:

X1 Collision Type(Head On)

X2	Collision Type(Rear End)
X3	Collision Type(Sidesweep)
X4	Traffic Control (Yes)
X5	Traffic Control (Other)
X6	Belted (Yes)
X7	Bike (Yes)
X8	Animal (Yes)
X9	Pedestrians (Yes)
X10	Area Type(Urban)
X11	Roadway Description (2-way divide)

In this section, looking at the results of the Multi-Level Multinomial Logistic regression results. The odds ratios for different variables and their potential impact on collision severity were obtained and analyzed. 'Collision Type (Head-On)' will be examined, with odds ratios of 1.9 for fatality, 3.6 for major injuries, and 1.2 for minor injuries. These numbers resonate with significance, revealing a heightened likelihood of both fatality and major injuries in head-on collisions. Minor injuries, on the other hand, display a relatively modest elevation. This output differentiation underscores the influence of collision type on the array of outcomes, as head-on collisions emerge as a potent driver of severe consequences (See Table 4.2).

Collision Type (Rear End):

For the Collision Type for Rear End, the odds ratios are 0.75 for fatality, 0.65 for major injuries, and 1.23 for minor injuries. A narrative of contrasts unfolds, showing a lowered likelihood of fatality and major injuries in rear-end collisions, perhaps attributed to their typically lower impact force or there would be cases of vehicle damage or property damage. However, minor injuries showcase a slight elevation. These odds ratios show the differential influence of collision types across varying injury levels (See Table 4.2).

Collision Type (Side sweep):

When looking at Collision-type side sweep, the odds ratios here are 0.68 for fatality, 0.38 for major injuries, and 0.47 for minor injuries. They paint a picture of reduced likelihood across the board, underscoring the potential protective effect associated with sideswipe collisions. The odds ratios resoundingly advocate for the relatively milder impact of side sweep collisions in mitigating the risk of severe outcomes.

Traffic Control:

The variable of 'Traffic Control' comes into the output, offering a glimpse into its influence. As contrasted with collisions without traffic control measures, those marked by traffic control bear odds ratios of 46 for fatality, 431 for major injuries, and 1 for minor injuries. These odds ratios shine a light on the substantial protective role that traffic control measures play, significantly reducing the likelihood of major injuries. Yet, it's noteworthy that the odds ratio of 1 for minor injuries suggests a similar likelihood, indicative of the balance maintained by these measures (See Table 4.2).

Traffic Control (Other):

Further looking at the narrative with the same Traffic control but with a different category, the category 'Traffic Control (Other)' comes to the fore. Relative to the baseline 'No' category, these collisions exhibit odds ratios of 18 for fatality, 345 for major injuries, and 0.84 for minor injuries. The unique odds ratios for each severity level underscore the differential impact of 'Traffic Control (Other)' on outcomes. These odds ratios echo the intricate interplay, reflecting the nuanced influence this variable imparts on each level of injury severity (See Table 4.2).

Belted (Yes):

The absence of seatbelt usage, embodied in 'Belted (No),' unfolds its story. Compared to the baseline 'Yes' category, this variable portrays odds ratios of 16 for fatality, 10 for major injuries, and 4 for minor injuries. These numbers are a stark reminder of the critical role seatbelt usage plays in averting severe outcomes. The odds ratios for fatality and major injuries emphasize the protective umbrella seatbelts provide, while minor injuries are significantly elevated in this context (See Table 4.2).

Bike:

The involvement of bicycles with the category 'Bike (Yes),' gives a unique hue to the narrative. Collisions featuring bicycles unfurl odds ratios of 13 for fatality, 67 for major injuries, and 41 for minor injuries, compared to the baseline. These odds ratios underline the heightened vulnerability associated with bicycle-involved collisions. The odds ratios for fatality and major injuries speak to the pivotal need for protective measures in these scenarios, while the substantial odds ratio for minor injuries underscores the potential for injury across the spectrum (See Table 4.2).

Animal:

The variable 'Animal (Yes)' introduces a novel facet. In contrast to the baseline 'No' category, these collisions present odds ratios of 5.7 for fatality, 0.56 for major injuries, and 1.5 for minor injuries. While the odds ratio of 5.7 for fatality should be interpreted cautiously due to the rarity of animal-related fatalities, the odds ratios for major and minor injuries indicate a unique impact. These odds ratios portray a measured elevation in risk, indicating the potential for animal-involved collisions to elevate the likelihood of injury (See Table 4.2).

Pedestrians:

As the labyrinth of collision dynamics is traversed, the presence of pedestrians emerges as a pivotal and impactful variable. Pedestrian-involved collisions project odds ratios of 354 for fatality, 436 for major injuries, and 258 for minor injuries when contrasted with the baseline 'No' category. The resounding message is unequivocal – pedestrian-involved collisions bear a significantly heightened risk across all levels of injury severity. These odds ratios accentuate the urgent need for focused efforts in enhancing pedestrian safety measures (See Table 4.2).

Area Type (Urban):

The 'Area Type (Urban)' variable gives a distinctive backdrop. Compared to its counterpart, 'Urban' scenarios embody odds ratios of 0.6 for fatality, 0.67 for major injuries, and 0.87 for minor injuries. These odds ratios spotlight the protective cloak urban areas offer, reducing the likelihood of severe outcomes. The odds ratios reflect the relatively safer environment urban settings provide across all levels of injury severity (See Table 4.2).

Roadway Description (2-way divide):

Lastly, 'Roadway Description (2-way divide)' takes its position. In contrast to the baseline 'One way' scenario, '2-way divide' unfolds odds ratios of 1.7 for fatality, 2.7 for major injuries, and 1.35 for minor injuries. These odds ratios unveil a nuanced panorama, signifying a heightened likelihood of severe outcomes in '2-way divide' contexts. Yet, minor injuries exhibit a relatively moderate elevation (See Table 4.2).

4.3 Models Comparison

In this section, the aim will be to look for the best model to predict collision severity, we worked on a task to compare two distinct models: the Multinomial Logistic Regression (MLR) and the Multi-Level Multinomial Logistic Regression (MLMLR) models. By examining key metrics such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), it can be determined which model offers a better fit for our data and research objectives.

AIC and BIC are like navigational tools that help us steer our models in the right direction. They consider the delicate balance between model performance and complexity. Lower AIC and BIC values are akin to a clearer path and a more efficient route towards understanding our data.

	Multinomial Logistic regression	Multi- Level Multinomial Logistic regression
AIC	427549	421118
BIC	428464	422365

When looking at the Multinomial Logistic Regression model the numbers are - AIC: 427549 and BIC: 428464. It presents us with an AIC value of approximately 427550 and a BIC value of around 428464. While interpreting these numbers, remember that smaller values are akin to a more streamlined model that fits the data well while avoiding unnecessary complexity. On the other hand, enters the Multi-Level Multinomial Logistic Regression model. This model unveils an AIC value of roughly 421118 and a BIC value of about 422365. Lower AIC and BIC indicate a better balance between fit and complexity, ultimately leading to clearer insights.

As this crossroads is navigated, it's crucial to remember that the lower the AIC and BIC, the better the model's performance in explaining our data while avoiding overfitting. By comparing these values, the model that best captures the intricate dance of collision severity can be discerned.

Based on our AIC and BIC analysis, the Multi-Level Multinomial Logistic Regression model emerges as the frontrunner. The MLMLR has a significantly lower AIC and DIC value (more than 10 units). With noticeably lower AIC and BIC values compared to the Multinomial Logistic Regression model, it demonstrates a superior balance between explaining the data and maintaining simplicity. This suggests that the MLMLR model likely captures underlying patterns and nuances in collision severity more effectively.

5. Machine Learning Modeling Results

This chapter will discuss the implementation and results of Machine Learning algorithms. The model which was used in the thesis is Random Forest Classifier.

5.1 Random Forest

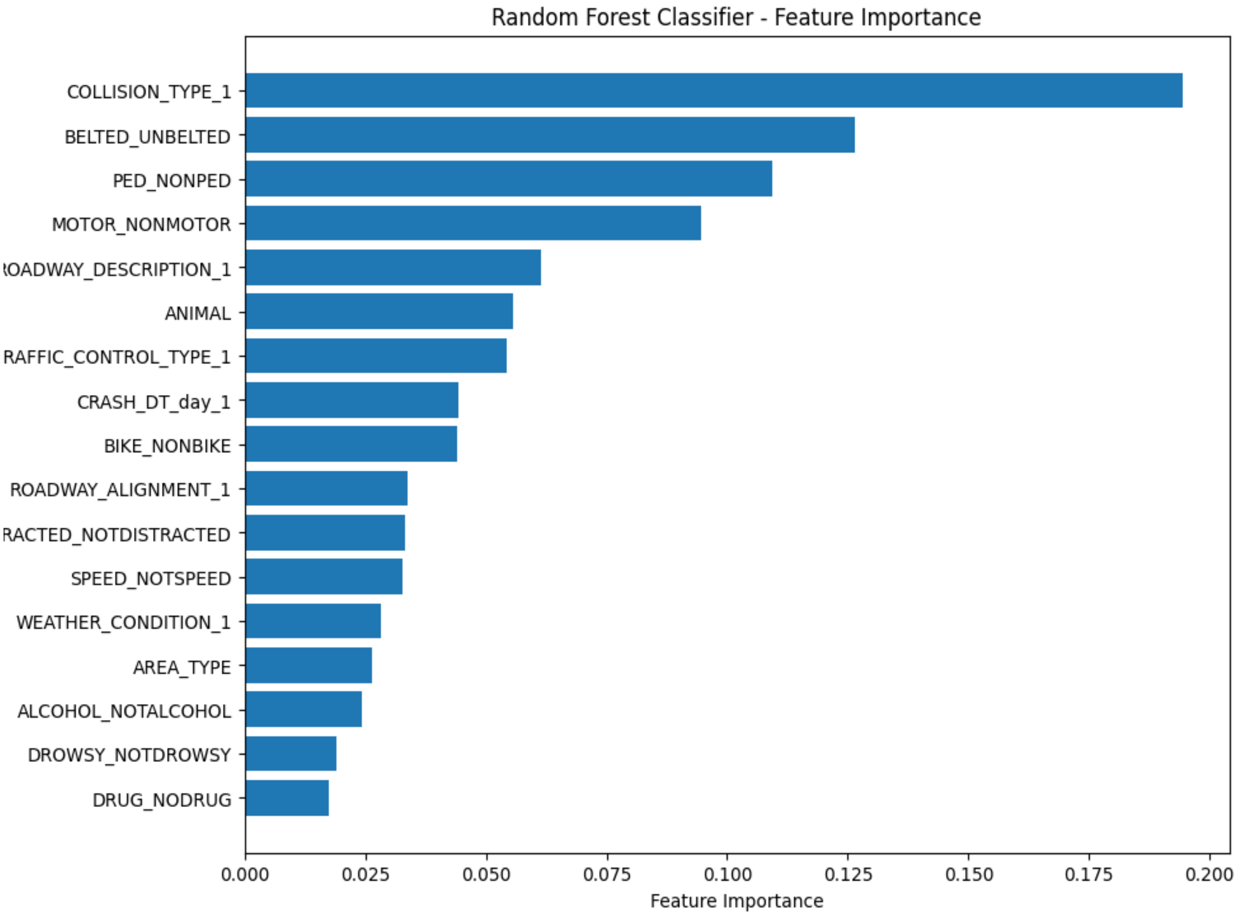
In understanding collision severity and its contributing factors, the analysis was done on a data-driven study using a Random Forest model. The model’s feature importance and accuracy provide us with valuable insights into the dynamics of collision severity prediction. A detailed analysis of the results and their implications is presented in this research study.

Figure 5.1 shows the visuals of the contribution of each variable in predicting the collision severity. The feature importance analysis is like finding or identifying a spotlight on the variables that hold the most influential factor in predicting collision severity. Each feature's important score reflects its contribution to the model's decision-making process. A hierarchy of influence can be discerned when looking at the list of features and their respective importance scores.

At the top of the list, "COLLISION_TYPE_1" is found with an importance score of approximately 19%: This underscores the significant impact of the collision type on predicting severity. Unsurprisingly, the way vehicles collide can greatly influence the outcome. Similarly, "BELTED_UNBELTED" and "PED_NONPED" follow closely, with importance scores of around 12% and 10%, respectively. This underscores the pivotal role of safety measures (seatbelts) and pedestrian involvement in determining the outcome of collisions.

Figure 5.1

Feature Importance for RF model



"ANIMAL," though appearing further down the list, still holds notable importance at around 8.9%. This highlights the significance of encounters with animals in road incidents, contributing significantly to the model's predictions. Other variables like "BIKE_NONBIKE," "ROADWAY_DESCRIPTION_1," and "TRFC_CTRL_STATUS_TYPE_1" hold moderate importance, underlining the varied dimensions that influence collision severity.

When assessing these important scores in the context of our thesis, a logical alignment can be found. The top features are those that intuitively impact collision severity—collision type, seatbelt usage, pedestrian involvement, and road description. These are factors we would naturally expect to influence the outcome. This reaffirms that our model is capturing meaningful patterns.

Table 5.1

Accuracy for our Random Forest Model

Accuracy	0.69		
AOC-ROC	0.93		
Macro Avg	0.44	0.31	0.32
Weighted Avg	0.63	0.69	0.61

When looking at the macro and weighted averages, how these metrics stack up overall can be observed. The macro average is around 32%, which indicates the general performance across all classes. The weighted average, considering the class distribution, lands at 61%, showcasing the model's overall performance considering class imbalances.

The accuracy of the model stands at 69%, reflecting the proportion of correctly predicted instances across all categories. While this is a decent overall accuracy, it's important to note that accuracy can sometimes be misleading when dealing with imbalanced datasets.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a vital metric in assessing the performance of binary classification models like the Random Forest. The ROC curve is a graphical representation of a model's ability to discriminate between positive and negative classes across different threshold values. The curve plots the True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) as the threshold for classification varies.

The AUC-ROC value, ranging from 0 to 1, summarizes the overall performance of the model. A higher AUC-ROC indicates better discrimination and model accuracy. An AUC-ROC of 0.5 suggests random guessing, while an AUC-ROC of 1 signifies perfect separation. In our results, the AUC-ROC of 0.937 indicates that the Random Forest model is effectively distinguishing between the two classes. This high value signifies strong predictive capability and a good balance between sensitivity and specificity. In practical terms, it means that when making predictions using this model, one can expect it to correctly rank positive instances higher than negative ones with a high probability.

The results can give us a conclusion that the Random Forest model performed impressively well with an AUC-ROC of 0.937. This indicates that the model has a high ability to differentiate between the classes, making it a promising choice for tasks requiring binary classification, such as disease diagnosis or fraud detection.

5.1.1 Simulating the RF Model

In this research study, a comprehensive exploration of predictive modeling using a Random Forest Classifier was conducted on the Collison Cleaned dataset.. The dataset, comprising variables such as COLLISION_TYPE_1, TRFC_CTRL_STATUS_TYPE_1, WEATHER_CONDITION_1, and others, aimed to predict the multi-class target variable, Crash_severity_3_level.

Upon building and fine-tuning the Random Forest model, the model's behavior under different circumstances was ventured into to extract insightful patterns. The top 8 significant variables, namely Belted, Animal, Pedestrians, Crash Day, Bike, Motor, and Collision type, were strategically modified.. Since these eight variables are the most important variables to predict the output and these simulations provided us with profound insights into the impact of these variables on the predicted outcomes, shedding light on the intricate relationships within the data.

BELTED:

In our analysis, the impact of the 'BELTED_UNBELTED' variable on predicted class labels for Collision Severity was simulated while keeping the other variables the same, using a Random Forest model. This variable represents whether a person was belted (0) or unbelted (1) during a collision. Our goal was to understand how this variable affects the severity of crashes, as indicated by different target class labels: Fatal, Major Injury, Minor Injury, PDO. When the results were visualized using a stacked bar plot, the distribution of these target class labels based on different values of the 'BELTED_UNBELTED' variable was observed.

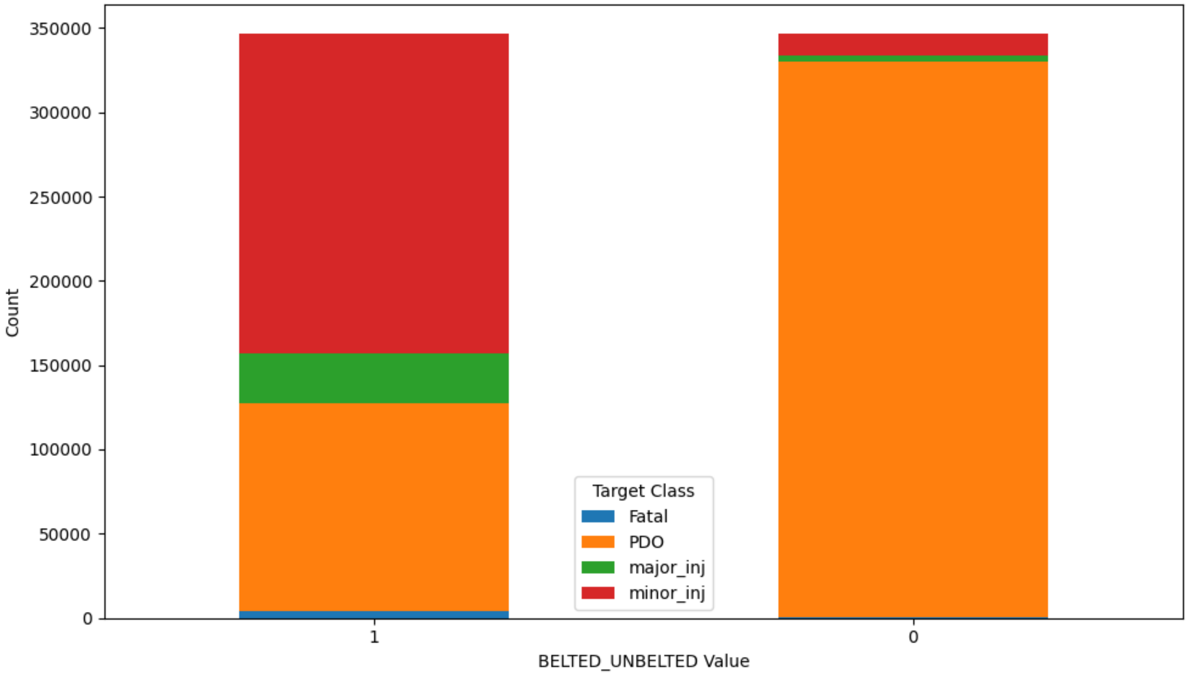
The stacked bar plot shows the distribution of target class labels for two scenarios: 'Belted' (0) and 'Unbelted' (1) individuals involved in collisions.

For 'Belted' Individuals (0):

The 'Fatal' category has a count of 348, indicating that 348 collisions involving belted individuals resulted in fatal outcomes. The 'PDO' (Property Damage Only) category has a count of 329,803, indicating that a large number of belted individuals had collisions resulting in property damage only. Similarly, it has counts of 3,279 and 13,050 for 'major_inj' (major injuries) and 'minor_inj' (minor injuries) categories, respectively.

Figure 5.2

Target Classes based on Simulating variable BELTED_UNBELTED



For 'Unbelted' Individuals (1):

The 'Fatal' category has a significantly higher count of 4,101, suggesting that collisions involving unbelted individuals were more likely to result in fatal outcomes. The 'PDO' category has a count of 123,295, indicating a substantial number of property damage-only collisions for unbelted individuals. The counts for 'major_inj' and 'minor_inj' categories are 29,756 and 189,328, respectively. This suggests a higher likelihood of major and minor injuries for unbelted individuals compared to belted ones

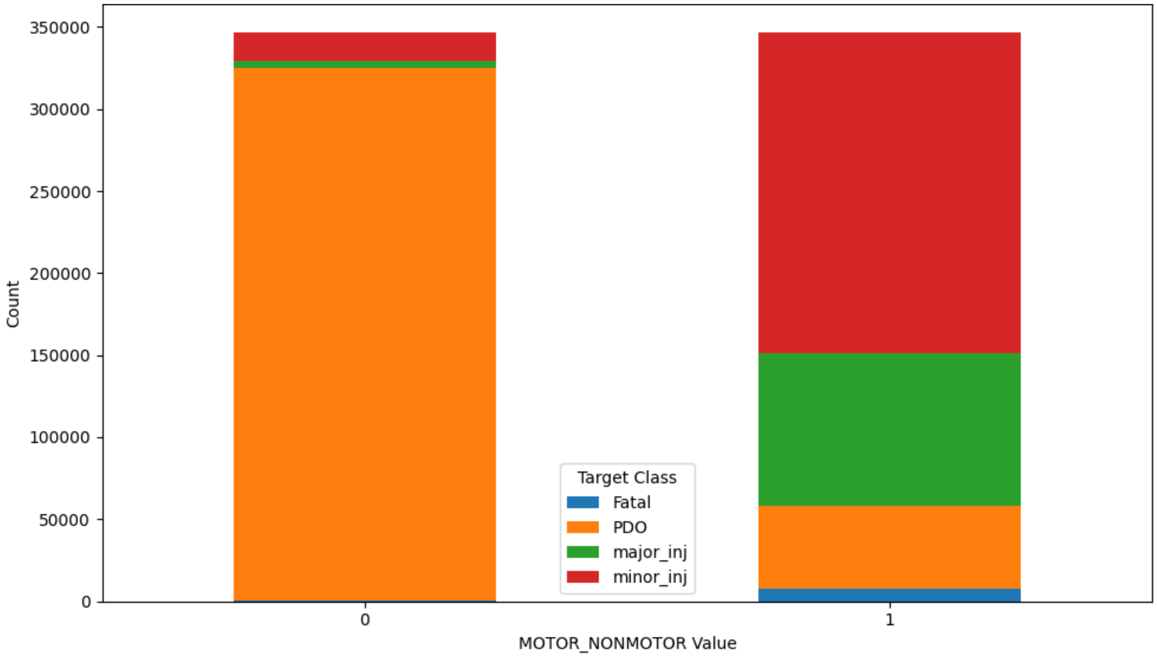
From the results and the visualization, it's evident that wearing seat belts plays a crucial role in reducing the severity of collision outcomes. Collisions involving belted individuals are more likely to result in property damage only, while unbelted individuals face a higher risk of fatal outcomes, major injuries, and minor injuries.

MOTOR:

The simulation focused on the 'MOTOR_NONMOTOR' variable's impact on predicted class labels can be seen in Figure 5.3. By altering the feature, the aim was to understand how different vehicle types influence crash severity. The output highlights key insights. When vehicles were identified as "MOTOR" (representing motorized vehicles), the majority of crashes resulted in Property Damage Only (PDO) incidents (324,504 cases) and minor injuries (17,095 cases). However, when considering "NONMOTOR" vehicles (non-motorized transport), there were fewer PDO incidents (50656) and minor injuries (195,073), with a notable increase in severe incidents. Fatalities were predominant (7452 cases) along with major injuries (93,299 cases).

Figure 5.3

Target Classes based on Simulating variable MOTOR_NONMOTOR



This indicates that non-motorized vehicles are more susceptible to severe outcomes. Such insights are crucial for designing safety measures for different vehicle types to mitigate severe incidents and ensure road user safety. The findings here emphasize the significance of vehicle type in predicting crash severity outcomes, which can inform policy and infrastructure improvements to address these disparities effectively.

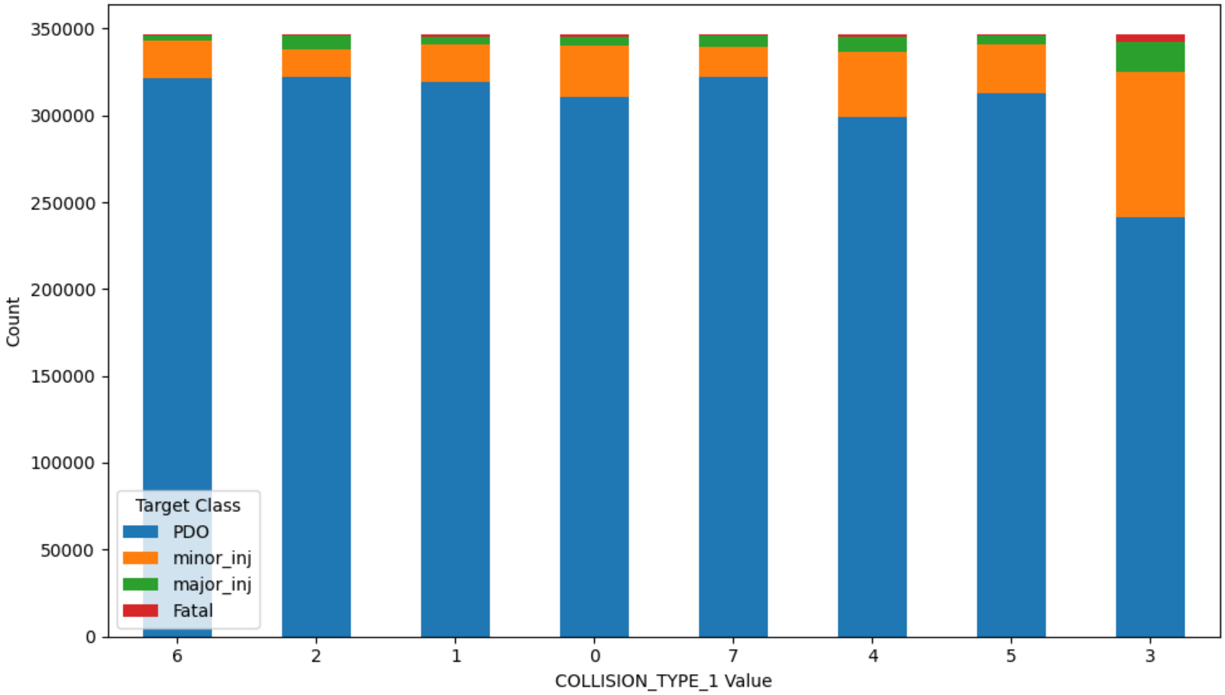
Collision Type:

In this part, the impact of Collision Type, such as Head-on, rear end, side sweep, etc., will be analyzed. The simulated impact of different collision types on predicted class labels provides valuable insights into the potential implications of varying collision scenarios on the severity of outcomes. and it's essential to interpret the results to uncover trends and implications for road safety.

The bar plot visualization showcases the distribution of predicted class labels across various collision types. Notably, collision type '6' appears to have the highest frequency of 'PDO' (Property Damage Only) outcomes, with a count of 321,198. Conversely, collision type '3' has the highest count of 'Fatal' outcomes at 4,371. This discrepancy suggests that collision type '6' might involve less severe accidents, while collision type '3' could correspond to more catastrophic incidents (See Figure 5.4).

Figure 5.4

Target Classes based on Simulating variable COLLISION_TYPE_1



Furthermore, analyzing the distribution of 'minor_inj' and 'major_inj' outcomes across collision types reveals interesting patterns. Collision type '3' again stands out with the highest counts in both categories, indicating a higher likelihood of serious injuries. On the other hand, collision type '2' displays the second-highest count in both categories, suggesting a consistent level of impact severity. The results also highlight variations in outcome severity within specific collision types. For instance, collision type '2' demonstrates a relatively balanced distribution across all four outcomes, while collision type '0' appears to have a higher count of 'Fatal' outcomes (See Figure 5.4).

Pedestrians:

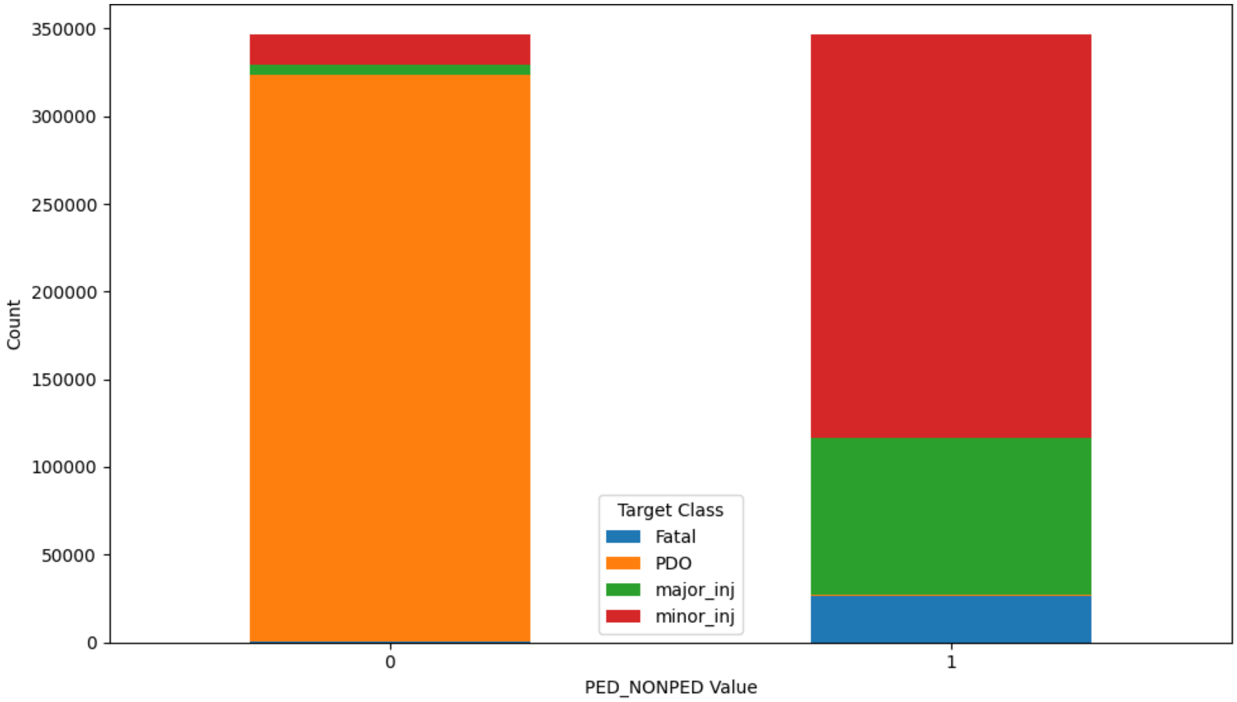
In the simulation that assesses the impact of different pedestrian involvement levels ('PED_NONPED') on predicted class labels, a comprehensive understanding of potential outcomes based on pedestrian presence or absence is unveiled. The visualization and numerical breakdown offer critical insights into the severity of predicted consequences in vehicular incidents involving pedestrians.

In figure 5.5 the bar plot presentation adeptly illustrates how varying levels of pedestrian involvement correlate with predicted class labels. Notably, instances with 'PED_NONPED' equal to '1' exhibit a substantially higher occurrence of 'Fatal' outcomes, totaling 26,567. This observation is noteworthy and suggests that incidents involving pedestrians often result in more severe consequences, underscoring the vulnerability of pedestrians on the road. Moreover, while 'minor_inj' outcomes still persist in the presence of pedestrians, they considerably outweigh 'major_inj' and 'Fatal' outcomes.

In contrast, when 'PED_NONPED' equals '0,' the incidence of 'Fatal' outcomes significantly decreases to 585. This trend implies that accidents in which pedestrians are not involved tend to result in fewer fatal incidents. Interestingly, 'PDO' (Property Damage Only) outcomes remain the most common in such scenarios, indicating relatively less severe accidents.

Figure 5.5

Target Classes based on Simulating variable PED_NONPED



Additionally, the count of 'PDO' outcomes is substantially higher across both levels of pedestrian involvement, suggesting that many accidents might involve only property damage, regardless of whether pedestrians are present or not.

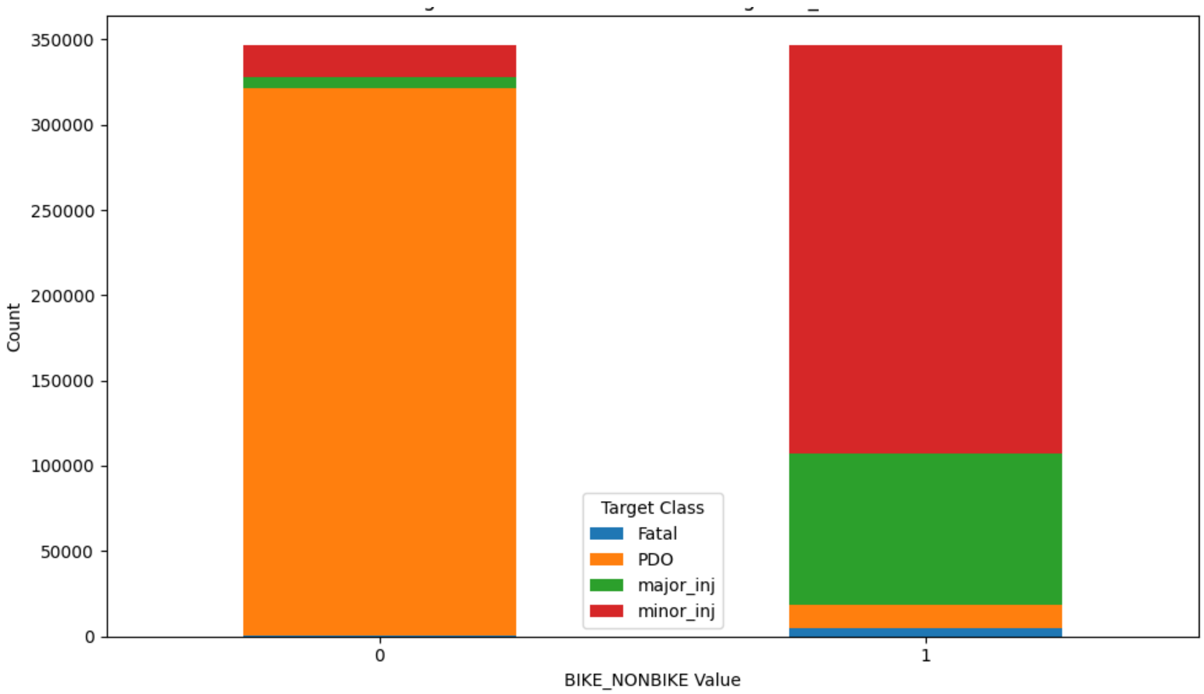
Bike:

The simulation exploring the impact of bicycle involvement ('BIKE_NONBIKE') on predicted class labels reveals compelling insights into the potential outcomes of accidents involving bicycles. The visual representation and quantitative breakdown offer a clear understanding of the severity of predicted consequences in incidents where bicycles are present on the road.

The stacked bar plot eloquently illustrates the distribution of predicted class labels based on the presence or absence of bicycles. Notably, instances with 'BIKE_NONBIKE' equal to '1' show a substantial count of 'minor_inj' outcomes, totaling 89,946. This observation suggests that accidents involving bicycles tend to result in a higher frequency of minor injuries. Furthermore, 'PDO' (Property Damage Only) outcomes are also prominent, signifying that accidents with bicycles often result in minimal physical harm but still involve property damage.

Figure 5.6

Target Classes based on Simulating variable BIKE-NONBIKE



Conversely, when 'BIKE_NONBIKE' equals '0,' the occurrences of 'minor_inj' outcomes decrease to 5,923. However, 'Fatal' outcomes in this scenario remain comparatively low at 732. This pattern indicates that accidents without bicycle involvement are less likely to result in fatalities, possibly due to lower impact forces. Moreover, 'PDO' outcomes persist as the most common outcome.

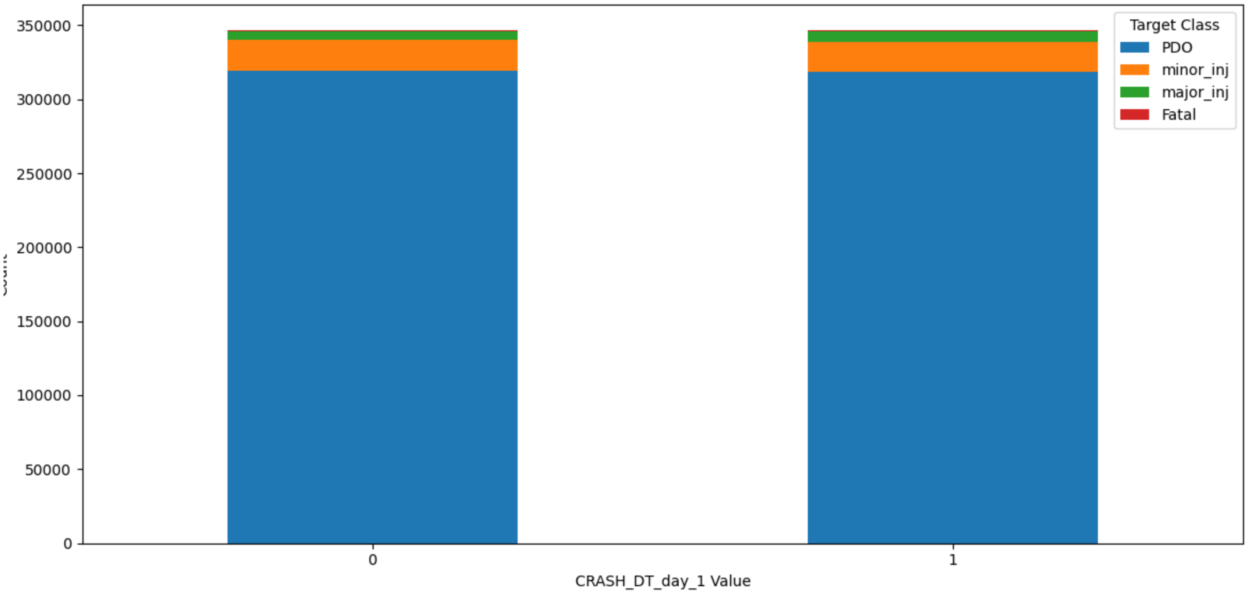
Interestingly, 'major_inj' outcomes occur regardless of bicycle involvement, emphasizing that severe injuries remain a possibility. Additionally, 'PDO' outcomes remain consistently high, regardless of bicycle presence or absence, further underscoring that many accidents result in only property damage.

Crash Day- Weekday:

The simulation investigating the influence of the day of the week on predicted accident outcomes ('CRASH_DT_day_1') provides valuable insights into the potential variations in accident severity based on different days. The visual representation and the accompanying numerical breakdown offer a comprehensive understanding of how accident outcomes differ across days of the week.

Figure 5.7

Target Classes based on Simulating variable CRASH DAY



The stacked bar plot effectively illustrates the distribution of predicted class labels for each day of the week. Notably, accidents occurring on both days ('CRASH_DT_day_1' = 0) and weekdays ('CRASH_DT_day_1' = 1) exhibit similar trends in predicted outcomes. The highest count of outcomes is 'PDO' (Property Damage Only), which signifies accidents with minimal physical harm but property damage. It's intriguing to observe that these outcomes are consistently prominent across all days (See Figure 5.7).

Interestingly, the counts of 'minor_inj' and 'major_inj' outcomes vary slightly between the two groups. Accidents on weekdays tend to have marginally higher counts of 'major_inj' outcomes, suggesting that accidents occurring during the workweek may involve higher impact forces. Conversely, accidents on both days have a slightly higher count of 'minor_inj' outcomes, indicating that minor injuries are more frequent in these cases.

Accuracy Assessment:

Accuracy serves as a benchmark to gauge the model's performance in classification tasks. In our case, the accuracy of approximately 68.4% signifies the proportion of correctly predicted collision severity levels. While this accuracy score provides a solid foundation, it also highlights the complexity of predicting such multifaceted events. Achieving higher accuracy would require a deep understanding of the intricate interplay between numerous variables influencing collision outcomes.

Interpreting the Results:

Our research takes a significant step forward with these results. The feature importance scores not only shed light on influential factors but also guide us toward areas that demand greater attention and further investigation. The dominance of collision type, seatbelt usage, and pedestrian involvement suggests that interventions aimed at reducing collision severity could potentially focus on these aspects.

Moreover, the accuracy score of our model signifies its capability to discern collision severity to a reasonable extent. This provides a solid foundation for policymakers, law enforcement, and stakeholders to develop informed strategies for road safety enhancements. However, it's essential to acknowledge the room for improvement. Collaboration between data scientists, traffic experts, and policymakers is crucial for refining the model's accuracy and ensuring its real-world applicability.

6. Conclusion

6.1 Summary of Findings

With this research study, the aim was to examine different factors contributing to traffic accident severities in North Virginia, US. In this research models that could not only address the issue but can also fit more than 2 categories in the output variables. The Multinomial Logistic Regression (MLR) and the Multi-level Multinomial Logistic Regression (MLMLR) models for our research. Upon analyzing the results from both the Multinomial Logistic Regression (MLR) and the Multi-level Multinomial Logistic Regression (MLMLR) models, distinct patterns and insights have obtained regarding the relationship between various factors and collision severity. The MLR model revealed several key factors that influence collision severity levels, including Collision Type, Traffic Control, Belt Usage, Bicycle Presence, Animal Incidents, and Pedestrian Involvement. These variables showed statistically significant coefficients, which indicated their significant impact on the outcome categories, such as Fatality, Major Injury, and Minor Injury, given their theoretical rationale of impacting collision severities. The results from this model showing the Odds ratio for fatality and major injury when Pedestrians and Bike are involved in accidents are 269, 28441 and 104, 4952 respectively, concluding that the collision where involvement of Pedestrians and Bike is present, those have the most severe impact leading to more Fatality and major injuries. Moreover, collisions where the driver is without the safety belts can lead to more of minor injuries. Collision occurs are traffic controls (e.g., signalized intersections) are likely to be more severe compared to collision occurs at regular road.

On the other hand, the MLMLR model, with its ability to account for hierarchical data structures, provided a more nuanced perspective. It reaffirmed the influence of factors like

Collision Type, Traffic Control, Belt Usage, Bicycle Presence, Animal Incidents, and Pedestrian Involvement. The results from the model show that accidents involving Pedestrians and Bike are more severe leading to Fatality and Major injuries. Moreover, additional variables like Area Type and Roadway Description were introduced, shedding light on their roles in influencing collision severity.

In comparison to the two models for predicting collision severity, it is evident that both the MLR and MLMLR models offer valuable insights into the relationship between predictor variables and collision severity outcomes. However, the MLMLR model brings an added advantage by accommodating the hierarchical nature of the data, yielding results that are potentially more robust and fine distinction.

6.2 Limitation and Future Work

Within the context of this research study, several limitations are acknowledged that warrant consideration and offer avenues for future investigations. It is imperative to recognize the inherent variability in crash record collection practices across different entities such as counties, governments, police agencies, and insurance companies. Since the dataset considered this study was from 9 districts of the North Virginia Transport Department, and every jurisdiction implemented its own rules and policies to report and collect information on crashes and collisions. This variation in data collection procedures may introduce inconsistencies in the available data, potentially leading to discrepancies in the included variables and their accuracy. To address this limitation, a concerted effort could be directed toward incorporating additional key and standardized variables that are universally relevant across jurisdictions. This expansion of variables would enhance the comprehensiveness of the model and contribute to its robustness, making it more adaptable to varying data collection practices.

Furthermore, the temporal scope of the dataset, spanning from 2019 to 2023, may be considered relatively limited in capturing the full spectrum of crash occurrences and associated influencing factors. Extending the dataset's timeline to encompass a more extensive range of years could lead to more substantial and statistically significant findings. A longer timeframe would enable the identification of trends and patterns that may emerge over time, thereby providing a more comprehensive understanding of the dynamics driving collision severity outcomes.

Also, by looking deeper into the relationships between the variables, we can uncover hidden patterns and nuanced insights that enhance the predictive accuracy of our models by analyzing the interdependency of variables. For instance, investigating scenarios where factors like alcohol impairment intersect with adverse weather conditions, such as slippery roads due to snow, could reveal critical risk factors that increase the likelihood of severe injuries. This multidimensional analysis promises a more comprehensive understanding of collision dynamics and paves the way for more effective accident prevention and response strategies. Future studies in this direction hold immense potential to contribute significantly to the field of road safety.

Sampling can also be considered as an issue or limitation. Usually, most of the collisions can be retrieved from Police reports, and there could be many instances when due to no injury or vehicle damage nothing was reported. However, a critical issue demanding attention is the potential under-reporting of crashes within the dataset. It is evident that not all crashes of significance are reported and subsequently collected. Back in 2009, a study was done by the National Highway Traffic Safety Administration to get insight into the realm of crash reporting, revealing intriguing insights. This investigation found that approximately 25% of minor injury crashes and fifty percent of no-injury crashes went unreported. This can play a huge role in making the result or outcomes biased. This inconsistency stands in stark contrast to the reporting rate observed for fatal crashes, which hovered around the 100 percent mark (National Highway Traffic Safety Administration, 2009; Blincoe et al., 2002). This under-reporting phenomenon could stem from a variety of factors, such as mild collisions not being deemed report-worthy or certain crashes occurring in remote areas with limited documentation.

This limitation underscores the importance of taking a cautious interpretation of the results and recognizing that they may not fully encapsulate the entire spectrum of crash severity incidents. Moreover, a specific challenge unique to the Canadian jurisdiction pertains to its reporting thresholds for collisions. Consequently, the crash records commonly referred to operate as outcome-based samples. This type of sampling pulls on the fact that the injury severities captured within police reports don't precisely give the genuine spectrum of crash incidents, primarily due to the underreporting of less severe injuries. This distinction casts a shadow on the accuracy of parameter estimates derived from outcome-based samples, potentially resulting in biased outcomes. Considering this, understanding the reporting criteria and their influence on data collection is pivotal for an accurate assessment of crash severity (Savolainen et al., 2011).

For future investigations, a more comprehensive exploration of these limitations could provide fruitful insights. Further research could focus on refining and expanding the model by incorporating additional covariates that address the variations in data collection practices among different entities. In conjunction, a cross-jurisdictional analysis could be pursued to assess the impact of reporting thresholds on the dataset's representativeness. Additionally, conducting a comparative study that encompasses multiple countries with varying reporting practices could offer a broader perspective on the relationship between collision severity and data collection methodologies.

In conclusion, while this research study has provided valuable insights into the relationships between predictor variables and collision severity outcomes, it is crucial to recognize the limitations inherent in the data and methodology. Future work could transcend these limitations by enhancing the model's comprehensiveness, extending the dataset's temporal scope, and investigating under-reporting and reporting thresholds. Also, to get a better insight into regions of Canada and specific to Vancouver, we can retrieve reliable data from the Police reports and then finalize 2-3 relevant models that can give us desired outcomes. Moreover, depending on the dataset, we can try to include more advanced Machine Learning models that can consider the broader picture including multiple variables, and result in better output. Addressing these aspects would contribute to the refinement and enrichment of the analytical approach, ultimately advancing our understanding of the intricate dynamics shaping collision severity outcomes.

References

1. Agyemang, W., Li, J., & Wu, C. (2019). Behavioral factors contributing to traffic crash severity. Journal of Transportation Safety & Security, 11(2), 151-165. https://doi.org/10.1080/19439962.2017.1389633
2. Alghnam, S., Towhari, J., Alkelya, M., Alsaif, A., Alrowaily, M., Alrabeeah, F., & Albabtain, I. (2019). The association between Mobile phone use and severe traffic injuries: a case-control study from Saudi Arabia. *International journal of environmental research and public health*, *16*(15), 2706.
3. Ahsan, H. M., & Sufian, A. A. (2014). Present condition and safety issues of non-motorized vehicles in Bangladesh. *Journal of Civil Engineering (IEB)*, *42*(1), 93-101
4. Anjuman, T., Siddiqui, C. K. A., Hasanat-E-Rabbi, S., & Hoque, M. M. (2013, March). Heavy vehicle driver involvement in road safety and multiple vehicle accidents in Bangladesh. In *Proceedings of the International Conference on Heavy Vehicles* (pp. 257-267). John Wiley & Sons, Hoboken, NJ
5. Azimian, A., Pyrialakou, V. D., Lavrenz, S., & Wen, S. (2021). Exploring the effects of area-level factors on traffic crash frequency by severity using multivariate space-time models. *Analytic Methods in Accident Research*, *31*, 100163.
6. Aziz, H. A., Ukkusuri, S. V., & Hasan, S. (2013). Exploring the determinants of pedestrian–vehicle crash severity in New York City. *Accident Analysis & Prevention*, *50*, 1298-1309.

7. Bahrololoom, S., Young, W., & Logan, D. (2017, November). A random parameter model of factors influencing bicycle fatal and serious injury crashes in Victoria, Australia. In *39th Australasian Transport Research Forum (ATRF), Auckland, New Zealand*.
8. BREIMAN, L., 1999, Random forests—random features. Technical Report 567, StatisticsDepartment, University of California, Berkeley, ftp://ftp.stat.berkeley.edu/pub/users/breiman
9. Canadian Institute of Actuaries. (2019). Motor Vehicle Collisions in Canada: Cost and Frequency Analysis. Retrieved from https://www.cia-ica.ca/docs/default-source/2019/219099e.pdf
10. Chen, P., & Shen, Q. (2019). Identifying high-risk built environments for severe bicycling injuries. *Journal of safety Research*, *68*, 1-7.
11. Ehsani, J. P., Bingham, C. R., Ionides, E., & Childers, D. (2014). The impact of Michigan's text messaging restriction on motor vehicle crashes. *Journal of Adolescent Health*, *54*(5), S68-S74.
12. Eluru, Naveen, Chandra R. Bhat, and David A. Hensher. "A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes." *Accident Analysis & Prevention* 40, no. 3 (2008): 1033-1054.
13. Hammad, H. M., Ashraf, M., Abbas, F., Bakhat, H. F., Qaisrani, S. A., Mubeen, M., ... & Awais, M. (2019). Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of Pakistan. *Environmental Science and Pollution Research*, *26*, 11674-11685.

14. Gamage, P., Karim, M. S., & Dozza, M. (2021). Understanding the influence of behavioral aspects in traffic collisions severities and its impact on economy. Transportation Research Interdisciplinary Perspectives, 10, 100367. https://doi.org/10.1016/j.trip.2021.100367
15. Hamer, M., Grzebieta, R., Williamson, A., & Olivier, J. (2021). Investigating the relationship between road design and serious injury crashes in Canada. Accident Analysis & Prevention, 153, 105975.
16. Jacobsen, M. R. (2013). Fuel economy and safety: The influences of vehicle class and driver behavior. *American Economic Journal: Applied Economics*, *5*(3), 1-26.
17. Kang, B. (2019). Identifying street design elements associated with vehicle-to-pedestrian collision reduction at intersections in New York City. *Accident Analysis & Prevention*, *122*, 308-317.
18. Lestina, D. C., Williams, A. F., Lund, A. K., Zador, P., & Kuhlmann, T. P. (1991). Motor vehicle crash injury patterns and the Virginia seat belt law. *Jama*, *265*(11), 1409-1413.
19. Musa, M. F., Hassan, S. A., & Mashros, N. (2020). The impact of roadway conditions towards accident severity on federal roads in Malaysia. *PLoS one*, *15*(7), e0235564.
20. Nguyen, H., Ivers, R. Q., Jan, S., Martiniuk, A. L., Li, Q., & Pham, C. (2013). The economic burden of road traffic injuries: evidence from a provincial general hospital in Vietnam. *Injury prevention*, *19*(2), 79-84.
21. Noland, R. B., & Quddus, M. A. (2004). A spatially disaggregate analysis of road casualties in England. Accident Analysis & Prevention, 36(6), 973-984. https://doi.org/10.1016/j.aap.2004.03.002

22. Osman, M., Paleti, R., Mishra, S., & Golias, M. M. (2016). Analysis of injury severity of large truck crashes in work zones. *Accident Analysis & Prevention*, *97*, 261-273.
23. Paleti, R., Eluru, N., & Bhat, C. R. (2010). Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis & Prevention*, *42*(6), 1839-1854.
24. Pang, T. Y., Cheung, Y. K., & Wong, S. C. (2019). Impact of cell phone use while driving on traffic safety in Canada: A review. Journal of Safety Research, 70, 129-136.
25. Pasha, M., Rifaat, S. M., Tay, R., & De Barros, A. (2016). Effects of street pattern, traffic, road infrastructure, socioeconomic and demographic characteristics on public transit ridership. *KSCE Journal of Civil Engineering*, *20*(3), 1017.
26. Quddus, M. (2015). Effects of geodemographic profiles of drivers on their injury severity from traffic crashes using multilevel mixed-effects ordered logit model. *Transportation research record*, *2514*(1), 149-157.
27. Rahimi, E., Shamshiripour, A., Samimi, A., & Mohammadian, A. K. (2020). Investigating the injury severity of single-vehicle truck crashes in a developing country. *Accident Analysis & Prevention*, *137*, 105444.
28. Robartes, E., & Chen, T. D. (2017). The effect of crash characteristics on cyclist injuries: An analysis of Virginia automobile-bicycle crash data. *Accident Analysis & Prevention*, *104*, 165-173
29. Rogeberg, O., & Elvik, R. (2016). The effects of cannabis intoxication on motor vehicle collision revisited and revised. *Addiction*, *111*(8), 1348-1359.

30. Rovšek, V., Batista, M., & Bogunović, B. (2017). Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametric classification tree. *Transport*, *32*(3), 272-281.

31. Safaei, B., Safaei, N., Masoud, A., & Seyedekrami, S. (2021). Weighing criteria and prioritizing strategies to reduce motorcycle-related injuries using combination of fuzzy TOPSIS and AHP methods. *Advances in transportation studies*, *54*.

32. Sapkota, D., Bista, B., & Adhikari, S. R. (2021). Economic costs associated with motorbike accidents in Kathmandu, Nepal. Journal of Health and Allied Sciences, 11(1), 1-5.

33. Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, *43*(5), 1666-1676.

34. Shrestha, R., Callaghan, J. P., & Taylor, E. (2017). The economic burden of alcohol-related collisions in Canada. Traffic Injury Prevention, 18(7), 724-730.

35. Tang, J., Gao, F., Liu, F., Han, C., & Lee, J. (2020). Spatial heterogeneity analysis of macro-level crashes using geographically weighted Poisson quantile regression. *Accident Analysis & Prevention*, *148*, 105833.

36. Tlaiss, H. A., & Baaj, M. H. (2020). Economic cost of road crashes: A systematic review. Traffic Injury Prevention, 21(1), 6-12. https://doi.org/10.1080/15389588.2019.1673778

37. Toran Pour, A., Moridpour, S., Tay, R., & Rajabifard, A. (2017). Neighborhood influences on vehicle-pedestrian crash severity. *Journal of urban health*, *94*, 855-868.

38. Transport Canada. (2018). Road safety in Canada 2018. Retrieved from https://www.tc.gc.ca/eng/motorvehiclesafety/tp-tp15145-1201.htm

39. Transport Canada. (2021). Canadian Motor Vehicle Traffic Collision Statistics: 2019. Retrieved from https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2019.html

40. WHO. (2021). Road traffic injuries. Retrieved from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

41. Wali, B., Khattak, A. J., & Xu, J. (2018). Contributory fault and level of personal injury to drivers involved in head-on collisions: Application of copula-based bivariate ordinal models. *Accident Analysis & Prevention*, *110*, 101-114.

42. Wang, C., Quddus, M. A., & Ison, S. G. (2013). The effect of traffic and road characteristics on road safety: A review and future research direction. *Safety science*, *57*, 264-275.

43. Zahabi, S.A.H., Strauss, J., Manaugh, K. and Miranda-Moreno, L.F., 2011. Estimating potential effect of speed limits, built environment, and other factors on severity of pedestrian and cyclist injuries in crashes. *Transportation research record*, *2247*(1), pp.81-90.

44. Zhang, G., Yau, K. K., & Zhang, X. (2014). Analyzing fault and severity in pedestrian–motor vehicle accidents in China. *Accident Analysis & Prevention*, *73*, 141-150.

45. Zhang, G., Yau, K. K., & Chen, G. (2013). Risk factors associated with traffic violations and accident severity in China. *Accident Analysis & Prevention*, *59*, 18-25.

46. Zhou, H., Yuan, C., Dong, N., Wong, S. C., & Xu, P. (2020). Severity of passenger injuries on public buses: A comparative analysis of collision injuries and non-collision injuries. *Journal of Safety Research*, *74*, 55-69.

47. Zou, Y., Zhang, Y., & Cheng, K. (2021). Exploring the impact of climate and extreme weather on fatal traffic accidents. *Sustainability*, *13*(1), 390.